Wiley Series in Probability and Statistics

SHAYLE R. SEARLE & MARVIN H. J. GRUBER

# LINEAR MODELS

SECOND EDITION

$$y=Xb+e$$


WILEY

# LINEAR MODELS

**WILEY SERIES IN PROBABILITY AND STATISTICS**

Established by *Walter A. Shewhart and Samuel S. Wilks*

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, Jozef L. Teugels*

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

# LINEAR MODELS

Second Edition

**SHAYLE R. SEARLE**
Cornell University, Ithaca, NY

**MARVIN H. J. GRUBER**
Rochester Institute of Technology, Rochester, NY

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada



*Library of Congress Cataloging-in-Publication Data is available.*

ISBN: 978-1-118-95283-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# CONTENTS

# PREFACE

I was both honored and humbled when, in November 2013, Stephen Quigley, then an associate publisher for John Wiley & Sons, now retired, asked me whether I would like to prepare a second edition of Searle's *Linear Models*. The first edition was my textbook when I studied linear models as a graduate student in statistics at the University of Rochester during the seventies. It has served me well as an important reference since then. I hope that this edition represents an improvement in the content, presentation, and timeliness of this well-respected classic. Indeed, *Linear Models* is a basic and very important tool for statistical analysis. The content and the level of this new edition is the same as the first edition with a number of additions and enhancements. There are also a few changes.

As pointed out in the first edition preface, the prerequisites for this book include a semester of matrix algebra and a year of statistical methods. In addition, knowledge of some of the topics in Gruber (2014) and Searle (2006) would be helpful.

The first edition had 11 chapters. The chapters in the new edition correspond to those in the first edition with a few changes and some additions. A short introductory chapter, Introduction and Overview is added at the beginning. This chapter gives a brief overview of what the entire book is about. Hopefully, this will give the reader some insight as to why some of the topics are taken up where they are. Chapters 1–10 are with additions and enhancements, the same as those of the first edition. Chapter 11, a list of formulae for estimating variance components in an unbalanced model is exactly as it was presented in the first edition. There are no changes in Chapter 11. This Chapter is available at the book's webpage www.wiley.com\go\Searle\LinearModels2E.

Here is how the content of Chapters 1–10 has been changed, added to, or enhanced.

In Chapter 1, the following topics have been added to the discussion of generalized inverses:

1. The singular value decomposition;
2. A representation of the Moore–Penrose inverse in terms of the singular value decomposition;
3. A representation of any generalized inverse in terms of the Moore–Penrose inverse;
4. A discussion of reflexive, least-square generalized, and minimum norm generalized inverses with an explanation of the relationships between them and the Moore–Penrose inverse.

The content of Chapter 2 is the same as that of the first edition with the omission of the section on singular normal distributions. Here, the reference is given to the first edition.

Chapter 3 has a number of additions and enhancements. Reviewers of the first edition claimed that the Gauss–Markov theorem was not discussed there. Actually, it was but not noted as such. I gave a formal statement and proof of this important result. I also gave an extension of the Gauss–Markov theorem to models where the parameters were random variables. This leads to a discussion of ridge-type estimators.

As was the case in the first edition, many of the numerical illustrations in Chapters 3–8 use hypothetical data. However, throughout the rest of the book, I have added some illustrative examples using real or simulated data collected from various sources. I have given SAS and R output for these data sets. In most cases, I did include the code. The advent of personal computers since the writing of the first edition makes this more relevant and easier to do than in 1971. When presenting hypothesis tests and confidence intervals, the notion of using $p$-values, as well as acceptance or rejection regions, was used. I made mention of how to calculate these values or obtain critical regions using graphing calculators like the TI 83 or 84. These enhancements were also made in the later chapters where appropriate.

Chapter 4 was pretty much the same as in the first edition with some changes in the exercises to make them more specific as opposed to being open-ended.

In addition to some of the enhancements mentioned for Chapter 3, Chapter 5 contains the following additional items:

1. Alternative definitions of estimable functions in terms of the singular value decomposition;
2. A formal statement and proof of the Gauss–Markov theorem for the non-full rank model using a Lagrange multiplier argument;
3. Specific examples using numbers in matrices of tests for estimability;
4. An example of how for hypothesis involving non-estimable functions using least-square estimators derived from different generalized inverses will yield different $F$-statistics.

In addition to the material of the first edition, Chapter 6 contains the following new items:

1. A few examples for the balanced case;
2. Some examples with either small "live" or simulated data sets;
3. A discussion of and examples of multiple comparisons, in particular Bonferonni and Scheffe simultaneous confidence intervals;
4. A discussion of the robustness of assumptions of normality, equal variances, and independent observations in analysis of variance;
5. Some non-parametric procedures for dealing with non-normal data;
6. A few examples illustrating the use of the computer packages SAS and R.

These items are also given for the two-way models that are considered in Chapter 7. In addition, an explanation of the difference between the Type I and Type III sum of squares in SAS is included. This is of particular importance for unbalanced data.

Chapter 8 presents three topics—missing values, analysis of covariance, and large-scale survey data. The second edition contains some numerical examples to illustrate why doing analysis considering covariates is important.

Chapter 9, in addition to the material in the first edition:

1. Illustrates "brute force" methods for computing expected mean squares in random and mixed models;
2. Clarifies and gives examples of tests of significance for variance components;
3. Presents and gives examples of the MINQUE, Bayes, and restricted Bayes estimator for estimating the variance components.

New in Chapter 10 are:

1. More discussion and examples of the MINQUE;
2. The connection between the maximum likelihood method and the best linear unbiased predictor.
3. Shrinkage methods for the estimation of variance components.

The references are listed after Chapter 10. They are all cited in the text. Many of them are new to the second edition and of course more recent. The format of the bibliography is the same as that of the first edition.

Chapter 11, the statistical tables from the first edition, and the answers to selected exercises are contained on the web page www.wiley.com\go\Searle\LinearModels2E. A solutions manual containing the solutions to all of the exercises is available to instructors using this book as a text for their course.

There are about 15% more exercises than in the first edition. Many of the exercises are those of the first edition, in some cases reworded to make them clearer and less open-ended.

The second edition contains more numerical examples and exercises than the first edition. Numerical exercises appear before the theoretical ones at the end of each chapter.

For the most part, notations are the same as those in the first edition. Letters in equations are italic. Vectors and matrices are boldfaced. With hopes of making reading easier, many of the longer sentences have been broken down to two or three simpler sentences. Sections containing material not in the first edition has been put in between the original sections where I thought it appropriate.

The method of numbering sections is the same as in the first edition using Arabic numbers for sections, lower case letters for sub-sections, and lower case roman numerals for sub-sub sections. Unlike the first edition, examples are numbered within each chapter as Example 1, Example 2, Example 3, etc., the numbering starting fresh in each new chapter. Examples end with □, formal proofs with ■. Formal definitions are in boxes.

I hope that I have created a second edition of this great work that is timely and reader-friendly. I appreciate any comments the readers may have about this.

A project like this never gets done without the help of other people. There were several members of the staff of John Wiley & Sons whom I would like to thank for help in various ways. My sincere thanks to Stephen H. Quigley, former Associate Publisher, for suggesting this project and for his helpful guidance during its early stages. I hope that he is enjoying his retirement. I would also like to express my gratitude to his successor Jon Gurstelle for his help in improving the timeliness of this work. I am grateful to Sari Friedman and Allison McGinniss and the production staff at Wiley for their work dealing with the final manuscript. In addition, I would like to thank the production editors Danielle LaCourciere of Wiley and Suresh Srinivasan of Aptara for the work on copyediting. Thanks are also due to Kathleen Pagliaro of Wiley for her work on the cover. The efforts of these people certainly made this a better book.

I would like to thank my teachers at the University of Rochester Reuben Gabriel, Govind Mudolkhar, and Poduri Rao for introducing me to linear models.

Special thanks go to Michal Barbosu, Head of the School of Mathematical Sciences at the Rochester Institute of Technology for helping to make SAS software available. I am grateful to my colleague Nathan Cahill and his graduate student Tommy Keane for help in the use of R statistical software.

I would like to dedicate this work to the memory of my parents Joseph and Adelaide Gruber. They were always there to encourage me during my growing up years and early adulthood.

I am grateful for the friendship of Frances Johnson and for the help and support she has given me over the years.

MARVIN H.J. GRUBER

*Rochester, NY*
*September 2016*

# PREFACE TO FIRST EDITION

This book describes general procedures of estimation and hypothesis testing for linear statistical models and shows their application for unbalanced data (i.e., unequal-subclass-numbers data) to certain specific models that often arise in research and survey work. In addition, three chapters are devoted to methods and results for estimating variance components, particularly from unbalanced data. Balanced data of the kind usually arising from designed experiments are treated very briefly, as just special cases of unbalanced data. Emphasis on unbalanced data is the backbone of the book, designed to assist those whose data cannot satisfy the strictures of carefully managed and well-designed experiments.

The title may suggest that this is an all-embracing treatment of linear models. This is not the case, for there is no detailed discussion of designed experiments. Moreover, the title is not *An Introduction to …*, because the book provides more than an introduction; nor is it *… with Applications,* because, although concerned with applications of general linear model theory to specific models, few applications in the form of-real-life data are used. Similarly, *… for Unbalanced Data* has also been excluded from the title because the book is not devoted exclusively to such data. Consequently the title *Linear Models* remains, and I believe it has brevity to recommend it.

My main objective is to describe linear model techniques for analyzing unbalanced data. In this sense the book is self-contained, based on prerequisites of a semester of matrix algebra and a year of statistical methods. The matrix algebra required is supplemented in Chapter 1, which deals with generalized inverse matrices and allied topics. The reader who wishes to pursue the mathematics in detail throughout the book should also have some knowledge of statistical theory. The requirements in this regard are supplemented by a summary review of distributions in Chapter 2,

extending to sections on the distribution of quadratic and bilinear forms and the singular multinormal distribution. There is no attempt to make this introductory material complete. It serves to provide the reader with foundations for developing results for the general linear model, and much of the detail of this and other chapters can be omitted by the reader whose training in mathematical statistics is sparse. However, he must know Theorems 1 through 3 of Chapter 2, for they are used extensively in succeeding chapters.

Chapter 3 deals with full-rank models. It begins with a simple explanation of regression (based on an example) and proceeds to multiple regression, giving a unified treatment for testing a general linear hypothesis. After dealing with various aspects of this hypothesis and special cases of it, the chapter ends with sections on reduced models and other related topics. Chapter 4 introduces models not of full rank by discussing regression on dummy (0, 1) variables and showing its equivalence to linear models. The results are well known to most statisticians, but not to many users of regression, especially those who are familiar with regression more in the form of computer output than as a statistical procedure. The chapter ends with a numerical example illustrating both the possibility of having many solutions to normal equations and the idea of estimable and non-estimable functions.

Chapter 5 deals with the non-full-rank model, utilizing generalized inverse matrices and giving a unified procedure for testing any testable linear hypothesis. Chapters 6 through 8 deal with specific cases of this model, giving many details for the analysis of unbalanced data. Within these chapters there is detailed discussion of certain topics that other books tend to ignore: restrictions on models and constraints on solutions (Sections 5.6 and 5.7); singular covariance matrices of the error terms (Section 5.8); orthogonal contrasts with unbalanced data (Section 5.5g); the hypotheses tested by $F$-statistics in the analysis of variance of unbalanced data (Sections 6.4f, 7.1g, and 7.2f); analysis of covariance for unbalanced data (Section 8.2); and approximate analyses for data that are only slightly unbalanced (Section 8.3). On these and other topics, I have tried to coordinate some ideas and make them readily accessible to students, rather than continuing to leave the literature relatively devoid of these topics or, at best, containing only scattered references to them. Statisticians concerned with analyzing unbalanced data on the basis of linear models have talked about the difficulties involved for many years but, probably because the problems are not easily resolved, little has been put in print about them. The time has arrived, I feel, for trying to fill this void. Readers may not always agree with what is said, indeed I may want to alter some things myself in due time but, meanwhile, if this book sets readers to thinking and writing further about these matters, I will feel justified. For example, there may be criticism of the discussion of $F$-statistics in parts of Chapters 6 through 8, where these statistics are used, not so much to test hypotheses of interest (as described in Chapter 5), but to specify what hypotheses are being tested by those $F$-statistics available in analysis of variance tables for unbalanced data. I believe it is important to understand what these hypotheses are, because they are not obvious analogs of the corresponding balanced data hypotheses and, in many cases, are relatively useless.

The many numerical illustrations and exercises in Chapters 3 through 8 use hypothetical data, designed with easy arithmetic in mind. This is because I agree with

C. C. Li (1964) who points out that we do not learn to solve quadratic equations by working with something like

$$683125x^2 + 1268.4071x - 213.69825 = 0$$

just because it occurs in real life. Learning to first solve $x^2 + 3x + 2 = 0$ is far more instructive. Whereas real-life examples are certainly motivating, they usually involve arithmetic that becomes as cumbersome and as difficult to follow as is the algebra it is meant to illustrate. Furthermore, if one is going to use real-life examples, they must come from a variety of sources in order to appeal to a wide audience, but the changing from one example to another as succeeding points of analysis are developed and illustrated brings an inevitable loss of continuity. No apology is made, therefore, for the artificiality of the numerical examples used, nor for repeated use of the same example in many places. The attributes of continuity and of relatively easy arithmetic more than compensate for the lack of reality by assuring that examples achieve their purpose, of illustrating the algebra.

Chapters 9 through 11 deal with variance components. The first part of Chapter 9 describes random models, distinguishing them from fixed models by a series of examples and using the concepts, rather than the details, of the examples to make the distinction. The second part of the chapter is the only occasion where balanced data are discussed in depth: not for specific models (designs) but in terms of procedures applicable to balanced data generally. Chapter 10 presents methods currently available for estimating variance components from unbalanced data, their properties, procedures, and difficulties. Parts of these two chapters draw heavily on Searle (1971). Finally, Chapter 11 catalogs results derived by applying to specific models some of the methods described in Chapter 10, gathering together the cumbersome algebraic expressions for variance component estimators and their variances in the 1-way, 2-way nested, and 2-way crossed classifications (random and mixed models), and others. Currently these results are scattered throughout the literature. The algebraic expressions are themselves so lengthy that there would be little advantage in giving numerical illustrations. Instead, extra space has been taken to typeset the algebraic expressions in as readable a manner as possible.

All chapters except the last have exercises, most of which are designed to encourage the student to reread the text and to practice and become thoroughly familiar with the techniques described. Statisticians, in their consulting capacity, are much like lawyers. They do not need to remember every technique exactly, but must know where to locate it when needed and be able to understand it once found. This is particularly so with the techniques of unbalanced data analysis, and so the exercises are directed towards impressing on the reader the methods and logic of establishing the techniques rather than the details of the results themselves. These can always be found when needed.

No computer programs are given. This would be an enormous task, with no certainty that such programs would be optimal when written and even less chance by the time they were published. While the need for good programs is obvious, I think that a statistics book is not the place yet for such programs. Computer programs

printed in books take on the aura of quality and authority, which, even if valid initially, soon becomes outmoded in today's fast-moving computer world.

The chapters are long, but self-contained and liberally sign-posted with sections, subsections, and sub-subsections—all with titles (see Contents).

My sincere thanks go to many people for helping with the book: the Institute of Statistics at Texas A. and M. University which provided me with facilities during a sabbatical leave (1968–1969) to do most of the initial writing; R. G. Cornell, N. R. Draper, and J. S. Hunter, the reviewers of the first draft who made many helpful suggestions; and my colleagues at Cornell who encouraged me to keep going. I also thank D. F. Cox, C. H. Goldsmith, A. Hedayat, R. R. Hocking, J. W. Rudan, D. L. Solomon, N. S. Urquhart, and D. L. Weeks for reading parts of the manuscript and suggesting valuable improvements. To John W. Rudan goes particular gratitude for generous help with proof reading. Grateful thanks also go to secretarial help at both Texas A. and M. and Cornell Universities, who eased the burden enormously.

S. R. SEARLE

*Ithaca, New York*
*October, 1970*

# ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

**www.wiley.com\go\Searle\LinearModels2E**

The website includes:

- Answers to selected exercises
- Chapter 11 from the first edition
- Statistical tables from the first edition

# INTRODUCTION AND OVERVIEW

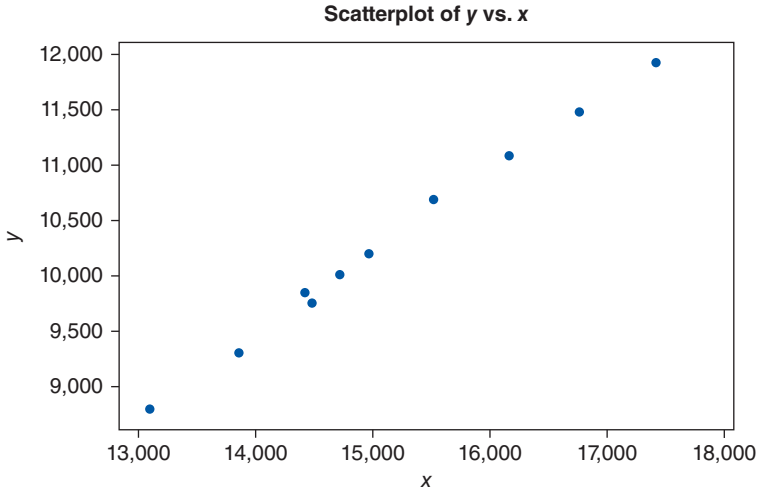
There are many practical real-world problems in many different disciplines where analysis using linear models is appropriate. We shall give several examples of such problems in this chapter as a motivation for the material in the succeeding chapters.

Suppose we consider personal consumption expenditures ($y$) in billions of dollars as a function of gross national product ($x$). Here are some data taken from the Economic Report of the President, 2015.

| Year | $x$ | $y$ |
|---|---|---|
| 2005 | 13,093.7 | 8,794.1 |
| 2006 | 13,855.9 | 9,304.0 |
| 2007 | 14,477.6 | 9,750.5 |
| 2008 | 14,718.6 | 10,013.6 |
| 2009 | 14,418.7 | 9,847.0 |
| 2010 | 14,964.4 | 10,202.2 |
| 2011 | 15,517.9 | 10,689.3 |
| 2012 | 16,163.2 | 11,083.1 |
| 2013 | 16,768.1 | 11,484.3 |
| 2014 | 17,420.7 | 11,928.4 |

Here is a scatterplot.

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.




The scatterplot suggests that a straight-line model $y = a + bx$ might be appropriate. The best fitting straight-line $y = -804.9 + 0.73412x$ accounts for 99.67% of the variation.

Suppose we have more independent variables, say $x_2$ (personal income in billions of dollars) and $x_3$ (the total number of employed people in the civilian labor force in thousands). The appropriate model might take the form (with $x_1$ the same as $x$ before)

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e,$$

where $e$ is an error term.

More generally, we will be considering models of the form

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

where $\mathbf{y}$ is an $N$-dimensional vector of observations, $\mathbf{X}$ is an $N \times (k + 1)$ matrix of the form $\begin{bmatrix} \mathbf{1}_N & \mathbf{X}_1 \end{bmatrix}$ where $\mathbf{1}_N$ is an $n$-dimensional vector of 1's and $\mathbf{X}_1$ is an $N \times k$ matrix of values of the independent variables, $\mathbf{b}$ is a $(k + 1)$-dimensional vector of regression parameters to be estimated, and $\mathbf{e}$ is an $N \times 1$ error vector. The estimators of $\mathbf{b}$ that we shall study most of the time will be least square estimators. These estimators minimize

$$F(\mathbf{b}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}).$$

We will show in Chapter 3 that, for full-rank matrices $\mathbf{X}$, they take the form

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

When $\mathbf{X}$ is not of full rank, the least-square estimators take the form

$$\hat{\mathbf{b}} = \mathbf{GX}'\mathbf{y},$$

where $\mathbf{G}$ is a generalized inverse $\mathbf{X}'\mathbf{X}$. We shall define generalized inverses and study their properties extensively in Chapter 1. We shall study the non-full-rank model in Chapter 5 and use the material presented there in the succeeding chapters.

In order to be able to make inferences about the regression parameters, for example, to find confidence intervals or perform hypothesis tests about them, we shall need the properties of the normal distribution and the distributions of functions of normal random variables. We shall study these distributions and their properties in Chapter 2.

Different forms of the $\mathbf{X}$ matrix will lead to different kinds of linear models for the solution of different kinds of problems. We shall now give some examples of these.

Suppose we wish to compare the life lengths of four different brands of light bulbs to see if there is a difference in their average life. For brands A, B, C, and D, we have life lengths

| A | B | C | D |
|---|---|---|---|
| 915 | 1011 | 989 | 1055 |
| 912 | 1001 | 979 | 1048 |
| 903 | 1003 | | 1061 |
| | 992 | | |

To represent the life lengths $y$ we use dummy variables. We have $x_1 = 1$ for an observation from brand A and $x_1 = 0$ for observations from brands B, C, and D. Likewise, $x_2 = 1$ for observations from brand B and $x_2 = 0$ for observations from brands A, C, and D. In a similar manner, $x_3 = 1$ for observations from brand C and $x_3 = 0$ for observations from brands A, B, and D. Also $x_4 = 1$ for observations from brand D and $x_4 = 0$ for observations from brands A, B, and C. The $y$'s are $y_{11}$, $y_{12}$, and $y_{13}$ for brand A; $y_{21}$, $y_{22}$, $y_{23}$, and $y_{24}$ for brand B; $y_{31}$ and $y_{32}$ for brand C; and $y_{41}$, $y_{42}$, and $y_{43}$ for brand D. If $\mu$ represents the intercept term we have that

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{41} \\ y_{42} \\ y_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + e$$

This is the familiar form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

where $\mathbf{b}' = \begin{bmatrix} \mu & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}$. The normal equations $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$ would be

$$\begin{bmatrix} 12 & 3 & 4 & 2 & 3 \\ 3 & 3 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 3 & 0 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{4.} \end{bmatrix}$$

or with numbers in non-matrix form

$$\begin{aligned} 12\hat{\mu} + 3\hat{\beta}_1 + 4\hat{\beta}_2 + 2\hat{\beta}_3 + 3\hat{\beta}_4 &= 11869 \\ 3\hat{\mu} + 3\hat{\beta}_1 &= 2730 \\ 4\hat{\mu} + 4\hat{\beta}_2 &= 4007 \\ 2\hat{\mu} + 2\hat{\beta}_3 &= 1968 \\ 3\hat{\mu} + 3\hat{\beta}_4 &= 3164 \end{aligned}$$

The $\mathbf{X}$ matrix here is of non-full rank so the system of equations has infinitely many solutions. To obtain solutions, we need to obtain a generalized inverse of $\mathbf{X}'\mathbf{X}$. There are infinitely many of them. They will be characterized in Chapters 1, 5, and 6. To determine which brands of light bulbs are different, we will have to conduct an analysis of variance to compare the mean life of the brands.

Actually there are two ways this experiment could be performed. One would be to just take specific brands. In this case, we would compare the mean life of the brands and make inferences about them. The other way would be to pick four brands of bulbs at random of each brand at random from many available bands. For this method, inferences would be about the variance components $\sigma^2_\beta$ because now the parameters would be random variables. We shall study methods of estimating and making inferences about variance components in Chapters 9 and 10.

We can also have mixed models where some of the effects are fixed effects and some of the effects are random effects. Such a model would take the form

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e},$$

where the $\beta's$ are fixed parameter values and the $\gamma$'s are random variables.

There are other situations where we would use a model of the form

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e},$$

where $\mathbf{X}$ is a matrix of 0's and 1's representing different factors and treatments and $\mathbf{Z}$ is numerical values of some quantity. The $\beta's$ and the $\gamma$'s are fixed parameter values.

For example, we could compare the weight loss of three groups of 10 people of three different reducing diets. The $\mathbf{X}$ matrix would consist of 0's and 1's using dummy variables. The $\mathbf{Z}$ matrix might contain information like the height and the weight of the 30 subjects before starting on the diets. Such variables are called covariates. In Chapter 8, we shall study the tool for analyzing such data, analysis of covariance.

Most of the time we will estimate parameters of linear models by least squares. However, there are situations where least-square estimators are not the best. This happens when the independent variables are highly correlated and the $\mathbf{X}'\mathbf{X}$ matrix is almost but not quite singular. Such data are called multicollinear and the least-square estimator may be very imprecise. One way to deal with such data is to use ridge regression. We shall discuss ridge-regression-type estimators in Chapter 3 and at other appropriate places in the text.

We begin by summarizing material on generalized inverses to be used in the later chapters.

# 1

# GENERALIZED INVERSE MATRICES

## 1. INTRODUCTION

Generalized inverse matrices are an important and useful mathematical tool for understanding certain aspects of the analysis procedures associated with linear models, especially the analysis of unbalanced data for non-full rank models. The analysis of unbalanced data and non-full rank models is of special importance and thus receives considerable attention in this book. Therefore, it is appropriate that we summarize the features of generalized inverses that are important to linear models. We will also discuss other useful and interesting results in matrix algebra.

We will frequently need to solve systems of equations of the form $\mathbf{Ax} = \mathbf{y}$ where $\mathbf{A}$ is an $m \times n$ matrix. When $m = n$ and $\mathbf{A}$ is nonsingular, the solution takes the form $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$.

For a consistent system of equations where $m$ may not equal $n$, or for square singular matrices, there exist matrices $\mathbf{G}$ where $\mathbf{x} = \mathbf{Gy}$. These matrices are generalized inverses.

**Example 1** Need for Generalized Inverses

Consider the system of equations

$$\begin{aligned} 5x_1 + 3x_2 + 2x_3 &= 50 \\ 3x_1 + 3x_2 &= 30 \\ 2x_1 + 2x_3 &= 20 \end{aligned}$$

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

or in matrix format

$$\begin{bmatrix} 5 & 3 & 2 \\ 3 & 3 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 50 \\ 30 \\ 20 \end{bmatrix}$$

Notice that the coefficient matrix is not of full rank. Indeed, the second and third rows add up to the first row. Solutions of this system include

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 50 \\ 30 \\ 20 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \\ 10 \end{bmatrix},$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{54} \begin{bmatrix} 5 & 1 & 4 \\ 1 & 11 & -10 \\ 4 & -10 & 14 \end{bmatrix} \begin{bmatrix} 50 \\ 30 \\ 20 \end{bmatrix} = \begin{bmatrix} \frac{20}{3} \\ \frac{10}{3} \\ \frac{10}{3} \end{bmatrix}$$

and infinitely many others. Each of the $3 \times 3$ matrices in the above solutions is generalized inverses. □

### a. Definition and Existence of a Generalized Inverse

In this book, we define a *generalized inverse* of a matrix $\mathbf{A}$ as any matrix $\mathbf{G}$ that satisfies the equation

$$\mathbf{AGA} = \mathbf{A}. \tag{1}$$

The reader may verify that the $3 \times 3$ matrices in the solutions to the system in Example 1 satisfies (1) and are thus, generalized inverses.

The name "generalized inverse" for matrices $\mathbf{G}$ defined by (1) is unfortunately not universally accepted. Names such as "conditional inverse," "pseudo inverse," and "g-inverse" are also to be found in the literature. Sometimes, these names refer to matrices defined as is $\mathbf{G}$ in (1) and sometimes to matrices defined as variants of $\mathbf{G}$. However, throughout this book, we use the name "generalized inverse" of $\mathbf{A}$ exclusively for any matrix $\mathbf{G}$ satisfying (1).

Notice that (1) does not define $\mathbf{G}$ as "the" generalized inverse of $\mathbf{A}$ but as "a" generalized inverse of $\mathbf{A}$. This is because $\mathbf{G}$, for a given matrix, $\mathbf{A}$ is not unique. As we shall show below there is an infinite number of matrices $\mathbf{G}$ that satisfy (1). Thus, we refer to the whole class of them as generalized inverses of $\mathbf{A}$.

Notice that in Example 1, we gave two generalized inverses of the coefficient matrix of the system of equations. Lots more could have been found.

There are many ways to find generalized inverses. We will give three here.

The first starts with the equivalent diagonal form of $\mathbf{A}$. If $\mathbf{A}$ has order $p \times q$, the reduction to this diagonal form can be written as

$$\mathbf{P}_{p\times p}\mathbf{A}_{p\times q}\mathbf{Q}_{q\times q} = \mathbf{\Delta}_{p\times q} \equiv \begin{bmatrix} \mathbf{D}_{r\times r} & \mathbf{0}_{r\times(q-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(q-r)} \end{bmatrix}$$

or more simply as

$$\mathbf{PAQ} = \mathbf{\Delta} = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2}$$

As usual, $\mathbf{P}$ and $\mathbf{Q}$ are products of elementary operators (see Searle, 1966, 2006, or Gruber, 2014). The matrix $\mathbf{A}$ has rank $r$ and $\mathbf{D}_r$ is a diagonal matrix of order $r$. In general, if $d_1, d_2, \ldots, d_r$ are the diagonal elements of any diagonal matrix $\mathbf{D}$, we will use the notation $\mathbf{D}\{d_i\}$for $\mathbf{D}_r$; that is,

$$\mathbf{D}_r \equiv \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ & & \ddots & \vdots \\ 0 & \cdots & 0 & d_r \end{bmatrix} \equiv \operatorname{diag}\{d_i\} = \mathbf{D}\{d_i\} \quad \text{for } i = 1, \ldots, r. \tag{3}$$

Furthermore, as in $\Delta$, the symbol $\mathbf{0}$ will represent null matrices with order being determined by the context on each occasion.

Derivation of $\mathbf{G}$ comes easily from $\Delta$. Analogous to $\Delta$, we define $\Delta^-$ (to be read $\Delta$ minus) as

$$\mathbf{\Delta}^- = \begin{bmatrix} \mathbf{D}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Then as shown below

$$\mathbf{G} = \mathbf{Q}\mathbf{\Delta}^-\mathbf{P} \tag{4}$$

satisfies (1) and is thus a generalized inverse. The generalized inverse $\mathbf{G}$ as given by (4) is not unique, because neither $\mathbf{P}$ nor $\mathbf{Q}$ by their definition is unique, neither is $\Delta$ or $\Delta^-$, and therefore $\mathbf{G} = \mathbf{Q}\Delta^-\mathbf{P}$ is not unique.

Before showing that $\mathbf{G}$ does satisfy (1), note from the definitions of $\Delta$ and $\Delta^-$ given above that

$$\mathbf{\Delta}\mathbf{\Delta}^-\mathbf{\Delta} = \mathbf{\Delta}. \tag{5}$$

Hence, by the definition implied in (1), we can say that $\Delta^-$ is a generalized inverse of $\Delta$. While this is an unimportant result in itself, it enables us to establish that $\mathbf{G}$,

as defined in (3), is indeed a generalized inverse of $\mathbf{A}$. To show this, observe that from (2),

$$\mathbf{A} = \mathbf{P}^{-1}\boldsymbol{\Delta}\mathbf{Q}^{-1}. \tag{6}$$

The inverses of $\mathbf{P}$ and $\mathbf{Q}$ exist because $\mathbf{P}$ and $\mathbf{Q}$ are products of elementary matrices and are, as a result, nonsingular. Then from (4), (5), and (6), we have,

$$\mathbf{AGA} = \mathbf{P}^{-1}\boldsymbol{\Delta}\mathbf{Q}^{-1}\mathbf{Q}\boldsymbol{\Delta}^{-}\mathbf{P}\mathbf{P}^{-1}\boldsymbol{\Delta}\mathbf{Q}^{-1} = \mathbf{P}^{-1}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-}\boldsymbol{\Delta}\mathbf{Q}^{-1} = \mathbf{P}^{-1}\boldsymbol{\Delta}\mathbf{Q}^{-1} = \mathbf{A}. \tag{7}$$

Thus, (1) is satisfied and $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$.

**Example 2** Obtaining a Generalized Inverse by Matrix Diagonalization
For

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 1 & 5 \\ 3 & 1 & 3 \end{bmatrix},$$

a diagonal form is obtained using

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 0 \\ -\frac{2}{3} & -\frac{1}{3} & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus,

$$\mathbf{PAQ} = \boldsymbol{\Delta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Delta}^{-} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

As a result,

$$\mathbf{G} = \mathbf{Q}\boldsymbol{\Delta}^{-}\mathbf{P} = \tfrac{1}{3}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The reader may verify that $\mathbf{AGA} = \mathbf{A}$. □

It is to be emphasized that generalized inverses exist for rectangular matrices as well as for square ones. This is evident from the formulation of $\boldsymbol{\Delta}_{p\times q}$. However, for $\mathbf{A}$ of order $p \times q$, we define $\boldsymbol{\Delta}^{-}$ as having order $q \times p$, the null matrices therein being

of appropriate order to make this so. As a result, the generalized inverse **G** has order $q \times p$.

**Example 3** Generalized Inverse for a Matrix That Is Not Square
Consider

$$\mathbf{B} = \begin{bmatrix} 4 & 1 & 2 & 0 \\ 1 & 1 & 5 & 15 \\ 3 & 1 & 3 & 5 \end{bmatrix}$$

the same **A** in the previous example with an additional column With **P** as given in Example 2 and **Q** now taken as

$$\mathbf{Q} = \begin{bmatrix} 1 & -1 & 1 & 5 \\ 0 & 1 & -6 & -20 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{PBQ} = \mathbf{\Delta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We then have

$$\mathbf{\Delta}^{-} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ so that } \mathbf{G} = \mathbf{Q\Delta}^{-}\mathbf{P} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{4}{3} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

□

### b. An Algorithm for Obtaining a Generalized Inverse

The algorithm is based on knowing or first finding the rank of the matrix. We present the algorithm first and then give a rationale for why it works. The algorithm goes as follows:

1. In **A** of rank $r$, find any non-singular minor of order $r$. Call it **M**.
2. Invert **M** and transpose the inverse to obtain $(\mathbf{M}^{-1})'$.
3. In **A,** replace each element of **M** by the corresponding element of $(\mathbf{M}^{-1})'$.
4. Replace all other elements of **A** by zero.
5. Transpose the resulting matrix.

The result is a generalized inverse of **A**. Observe that different choices of the minor of rank $r$ will give different generalized inverses of **A**.

**Example 4** Computing a Generalized Inverse using the Algorithm
Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ 0 & 1 & -3 & -2 \end{bmatrix}.$$

The reader may verify that all of the $3 \times 3$ sub-matrices of $\mathbf{A}$ have determinant zero while the $2 \times 2$ sub-matrices have non-zero determinants. Thus, $\mathbf{A}$ has rank 2. Consider

$$\mathbf{M} = \begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix}.$$

Then

$$\mathbf{M}^{-1} = \begin{bmatrix} 7 & -2 \\ -3 & 1 \end{bmatrix}$$

and

$$(\mathbf{M}^{-1})' = \begin{bmatrix} 7 & -3 \\ -2 & 1 \end{bmatrix}.$$

Now write the matrix

$$\mathbf{H} = \begin{bmatrix} 7 & -3 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Then the generalized inverse

$$\mathbf{G} = \mathbf{H}' = \begin{bmatrix} 7 & -2 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

By a similar process, taking

$$\mathbf{M} = \begin{bmatrix} 12 & 4 \\ -3 & -2 \end{bmatrix},$$

another generalized inverse of $\mathbf{A}$ is

$$\tilde{\mathbf{G}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{6} & \frac{1}{3} \\ 0 & -\frac{1}{4} & -1 \end{bmatrix}.$$

The reader may, if he/she wishes, construct other generalized inverses using $2 \times 2$ sub-matrices with non-zero determinant. □

We now present the rationale for the algorithm. Suppose $\mathbf{A}$ can be partitioned in such a way that its leading $r \times r$ minor is non-singular, that is,

$$\mathbf{A}_{p\times q} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is $r \times r$ of rank $r$. Then a generalized inverse of $\mathbf{A}$ is

$$\mathbf{G}_{q\times p} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

where the null matrices are of appropriate order to make $\mathbf{G}$ a $q \times p$ matrix. To see that $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$, note that

$$\mathbf{AGA} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix}.$$

Now since $\mathbf{A}$ is of rank $r$, the rows are linearly dependent. Thus, for some matrix $\mathbf{K}$ $[\,\mathbf{A}_{21} \quad \mathbf{A}_{22}\,] = \mathbf{K}[\,\mathbf{A}_{11} \quad \mathbf{A}_{12}\,]$. Specifically $\mathbf{K} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$ and so $\mathbf{A}_{22} = \mathbf{K}\mathbf{A}_{12} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. Hence, $\mathbf{AGA} = \mathbf{A}$ and $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$.

There is no need for the non-singular minor to be in the leading position. Let $\mathbf{R}$ and $\mathbf{S}$ represent the elementary row and column operations, respectively, to bring it to the leading position. Then $\mathbf{R}$ and $\mathbf{S}$ are products of elementary operators with

$$\mathbf{RAS} = \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \tag{8}$$

where $\mathbf{B}_{11}$ is non-singular of order $r$. Then

$$\mathbf{F} = \begin{bmatrix} \mathbf{B}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

is a generalized inverse of $\mathbf{B}$ and $\mathbf{G}_{q\times p} = \mathbf{SFR}$ is a generalized inverse of A. From (8),

$$\mathbf{A} = \mathbf{R}^{-1}\mathbf{B}\mathbf{S}^{-1}.$$

Then

$$\mathbf{AGA} = \mathbf{R}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{SFRR}^{-1}\mathbf{B}\mathbf{S}^{-1} = \mathbf{R}^{-1}\mathbf{BFBS}^{-1} = \mathbf{R}^{-1}\mathbf{B}\mathbf{S}^{-1} = \mathbf{A}.$$

Now $\mathbf{R}$ and $\mathbf{S}$ are products of elementary operators that exchange rows and columns. Such matrices are identity matrices with rows and columns interchanged. Such matrices are known as *permutation matrices* and are orthogonal. Thus, we have that $\mathbf{R} = \mathbf{I}$ with its rows in a different sequence, a permutation matrix and $\mathbf{R}'\mathbf{R} = \mathbf{I}$. The same is true for S and so from (8), we have that

$$\mathbf{A} = \mathbf{R}'\mathbf{B}\mathbf{S}' = \mathbf{R}'\begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\mathbf{S}'. \tag{9}$$

As far as $\mathbf{B}_{11}$ is concerned, the product in (9) represents the operations of returning the elements of $\mathbf{B}_{11}$ to their original position in $\mathbf{A}$. Now consider $\mathbf{G}$. We have

$$\mathbf{G} = \mathbf{SFR} = (\mathbf{R}'\mathbf{F}'\mathbf{S}')' = \left\{\mathbf{R}'\begin{bmatrix}(\mathbf{B}_{11}^{-1})' & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{S}'\right\}$$

In this, analogous to the form of $\mathbf{A} = \mathbf{R}'\mathbf{B}\mathbf{S}'$ the product involving $\mathbf{R}'$ and $\mathbf{S}'$ in $\mathbf{G}'$ represents putting the elements of $(\mathbf{B}_{11}^{-1})'$ into the corresponding positions of $\mathbf{G}'$ that the elements of $\mathbf{B}_{11}$ occupied in $\mathbf{A}$. This is what motivates the algorithm.

### c. Obtaining Generalized Inverses Using the Singular Value Decomposition (SVD)

Let $\mathbf{A}$ be a matrix of rank $r$. Let $\mathbf{\Lambda}$ be $r \times r$ the diagonal matrix of non-zero eigenvalues of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ ordered from largest to smallest. The non-zero eigenvalues of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ are the same (see p. 110 of Gruber (2014) for a proof). Then the decomposition of

$$\mathbf{A} = \begin{bmatrix}\mathbf{S}' & \mathbf{T}'\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}^{1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}'\\ \mathbf{V}'\end{bmatrix} = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}', \tag{10}$$

where $[\,\mathbf{S}' \quad \mathbf{T}'\,]$ and $[\,\mathbf{U} \quad \mathbf{V}\,]$ are orthogonal matrices, is the singular value decomposition (SVD). The existence of this decomposition is established in Gruber (2014) following Stewart (1963, p. 126). Observe that $\mathbf{S}'\mathbf{S} + \mathbf{T}'\mathbf{T} = \mathbf{I}, \mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}' = \mathbf{I}, \mathbf{S}\mathbf{S}' = \mathbf{I}, \mathbf{T}\mathbf{T}' = \mathbf{I}, \mathbf{S}'\mathbf{T} = \mathbf{0}, \mathbf{T}'\mathbf{S} = \mathbf{0}, \mathbf{U}\mathbf{U}' = \mathbf{I}, \mathbf{U}'\mathbf{V} = \mathbf{0},$ and $\mathbf{V}'\mathbf{U} = \mathbf{0}$. Furthermore, $\mathbf{A}'\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ and $\mathbf{A}\mathbf{A}' = \mathbf{S}'\mathbf{\Lambda}\mathbf{S}$. A generalized inverse of $\mathbf{A}$ then takes the form

$$\mathbf{G} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{S}. \tag{11}$$

Indeed, $\mathbf{AGA} = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}'\mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{S}\mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{A}$.

**Example 5** Finding a Generalized Inverse using the Singular Value Decomposition
Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Then,

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{A}' = \begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \end{bmatrix}.$$

To find the eigenvalues of $\mathbf{A}'\mathbf{A}$ solve the equation

$$\det \begin{bmatrix} 4-\lambda & 2 & 2 \\ 2 & 2-\lambda & 0 \\ 2 & 0 & 2-\lambda \end{bmatrix} = 0$$

or

$$\lambda^3 - 8\lambda^2 + 12\lambda = \lambda(\lambda - 6)(\lambda - 2) = 0$$

to get the eigenvalues $\lambda = 6, 2, 0$. Finding the eigenvectors by solving the systems of equations

$$\begin{matrix} -2x_1 + 2x_2 + 2x_3 = 0 \\ 2x_1 - 4x_2 = 0 \\ 2x_1 - 4x_3 = 0 \end{matrix} \qquad \begin{matrix} 2x_1 + 2x_2 + 2x_3 = 0 \\ 2x_1 = 0 \end{matrix} \qquad \begin{matrix} 4x_1 + 2x_2 + 2x_3 = 0 \\ 2x_1 + 2x_2 = 0 \\ 2x_1 + 2x_3 = 0 \end{matrix}$$

yields a matrix of normalized eigenvectors of $\mathbf{A}'\mathbf{A}$,

$$[\,\mathbf{U} \quad \mathbf{V}\,] = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix}.$$

By a similar process, the reader may show that the eigenvalues of $\mathbf{AA}'$ are $\lambda = 6, 2, 0, 0$ and that the matrix of eigenvectors is

$$\begin{bmatrix} \mathbf{S}' & \mathbf{T}' \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

Then the singular value decomposition of

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

and, as a result, the generalized inverse

$$\mathbf{G} = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

□

These derivations of a generalized inverse matrix $\mathbf{G}$ are by no means the only ways such a matrix can be computed. For matrices of small order, they can be satisfactory, but for those of large order that might occur in the analysis of "big data," other methods might be preferred. Some of these are discussed subsequently. Most methods involve, of course, the same kind of numerical problems as are incurred in calculating the regular inverse $\mathbf{A}^{-1}$ of a non-singular matrix $\mathbf{A}$. Despite this, the generalized inverse has importance because of its general application to non-square matrices and to square singular matrices. In the special case that $\mathbf{A}$ is non-singular, $\mathbf{G} = \mathbf{A}^{-1}$ as one would expect, and in this case, $\mathbf{G}$ is unique.

The fact that $\mathbf{A}$ has a generalized inverse even when it is singular or rectangular has particular application in the problem of solving equations, for example, of solving $\mathbf{Ax} = \mathbf{y}$ for $\mathbf{x}$ when $\mathbf{A}$ is singular or rectangular. In situations of this nature, the use of a generalized inverse $\mathbf{G}$, as we shall see, leads very directly to a solution in the form $\mathbf{x} = \mathbf{Gy}$. This is of great importance in the study of linear models where such equations arise quite frequently. For example, when we can write a linear model as $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, finding the least square estimator for estimating $\mathbf{b}$ leads to equations $\mathbf{X'X}\hat{\mathbf{b}} = \mathbf{X'y}$ where the matrix $\mathbf{X'X}$ is singular. Hence, we cannot write the solution as $(\mathbf{X'X})^{-1}\mathbf{X'y}$. However, using a generalized inverse $\mathbf{G}$ of $\mathbf{X'X}$, we can obtain the solution directly in the form $\mathbf{GX'y}$ and study its properties.

For linear models, the use of generalized inverse matrices in solving linear equations is the application of prime interest. We now outline the resulting procedures. Following this, we discuss some general properties of generalized inverses.

## 2. SOLVING LINEAR EQUATIONS

### a. Consistent Equations

A convenient starting point from which to develop the solution of linear equations using a generalized inverse is the definition of consistent equations.

**Definition 1** The linear equations $\mathbf{Ax} = \mathbf{y}$ are defined as being consistent if any linear relationships existing among the rows of $\mathbf{A}$ also exist among the corresponding elements of $\mathbf{y}$. In other words, $\mathbf{t'A} = 0$ if and only if $\mathbf{t'y} = 0$ for any vector $\mathbf{t}$.

As a simple example, the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \end{bmatrix}$$

are consistent. The second row of the matrix on the left-hand side of the system is the first row multiplied and on the right-hand side, of course $21 = 7(3)$. On the other hand, the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 24 \end{bmatrix}$$

are inconsistent. The linear relationship between the rows of the matrix on the left-hand side of the system does not hold true between 7 and 24. Moreover, you can write out the two equations and show that $3 = 0$.

The formal definition of consistent equations does not demand that linear relationships exist among the rows of $\mathbf{A}$. However, if they do, then the definition does require that the same relationships also exist among the corresponding elements of $\mathbf{y}$ for the equations to be consistent. For example, when $\mathbf{A}$ is non-singular, the equations

$\mathbf{Ax} = \mathbf{y}$ are always consistent. There are no linear relationships among the rows of $\mathbf{A}$ and therefore none that the elements of $\mathbf{y}$ must satisfy.

The importance of consistency lies in the following theorem. Linear equations can be solved only if they are consistent. See, for example, Section 6.2 of Searle (1966) or Section 7.2 of Searle and Hausman (1970) for a proof. Since only consistent equations can be solved, discussion of a procedure for solving linear equations is hereafter confined to equations that are consistent. The procedure is described in Theorems 1 and 2 in Section 2b. Theorems 3–6 in Section 2c deal with the properties of these solutions.

## b. Obtaining Solutions

The link between a generalized inverse of the matrix $\mathbf{A}$ and consistent equations $\mathbf{Ax} = \mathbf{y}$ is set out in the following theorem adapted from C. R. Rao (1962).

**Theorem 1** Consistent equations $\mathbf{Ax} = \mathbf{y}$ have a solution $\mathbf{x} = \mathbf{Gy}$ if and only if $\mathbf{AGA} = \mathbf{A}$.

*Proof.* If the equations $\mathbf{Ax} = \mathbf{y}$ are consistent and have $\mathbf{x} = \mathbf{Gy}$ as a solution, write $\mathbf{a}_j$ for the $j$th column of $\mathbf{A}$ and consider the equations $\mathbf{Ax} = \mathbf{a}_j$. They have a solution. It is the null vector with its $j$th element set equal to unity. Therefore, the equations $\mathbf{Ax} = \mathbf{a}_j$ are consistent. Furthermore, since consistent equations $\mathbf{Ax} = \mathbf{y}$ have a solution $\mathbf{x} = \mathbf{Gy}$, it follows that consistent equations $\mathbf{Ax} = \mathbf{a}_j$ have a solution $\mathbf{x} = \mathbf{Ga}_j$. Therefore,

$\mathbf{AGa}_j = \mathbf{a}_j$ and this is true for all values of $j$, that is, for all columns of $\mathbf{A}$. Hence, $\mathbf{AGA} = \mathbf{A}$.

Conversely, if $\mathbf{AGA} = \mathbf{A}$ then $\mathbf{AGAx} = \mathbf{Ax}$, and when $\mathbf{Ax} = \mathbf{y}$ substitution gives $\mathbf{A}(\mathbf{Gy}) = \mathbf{y}$. Hence, $\mathbf{x} = \mathbf{Gy}$ is a solution of $\mathbf{Ax} = \mathbf{y}$ and the theorem is proved. ■

Theorem 1 indicates how a solution to consistent equations may be obtained. Find any generalized inverse of $\mathbf{A}$, $\mathbf{G}$, and then $\mathbf{Gy}$ is a solution. However, this solution is not unique. There are, indeed, many solutions whenever $\mathbf{A}$ is anything but a square, non-singular matrix. These are characterized in Theorem 2 and 3.

**Theorem 2** If $\mathbf{A}$ has $q$ columns and $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$, then the consistent equations $\mathbf{Ax} = \mathbf{y}$ have the solution

$$\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{GA} - \mathbf{I})\mathbf{z} \tag{12}$$

where $\mathbf{z}$ is any arbitrary vector of order $q$.

*Proof.* Since $\mathbf{AGA} = \mathbf{A}$, $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{AGy} + (\mathbf{AGA} - \mathbf{A})\mathbf{z} = \mathbf{AGy} = \mathbf{y}$, by Theorem 1. ■

There are as many solutions to (12) as there are choices of $\mathbf{z}$ and $\mathbf{G}$. Thus, the equation $\mathbf{Ax} = \mathbf{y}$ has infinitely many solutions of the form (12).

**Example 6** Different Solutions to Ax = y for a particular A

Consider the equations **Ax = y as**

$$\begin{bmatrix} 5 & 3 & 1 & -4 \\ 8 & 5 & 2 & 3 \\ 21 & 13 & 5 & 2 \\ 3 & 2 & 1 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 22 \\ 2 \end{bmatrix}, \tag{13}$$

so defining **A**, **x**, and **y**. Using the algorithm developed in Section 1b with the $2 \times 2$ minor in the upper left-hand corner of **A**, it will be found that

$$\mathbf{G} = \begin{bmatrix} 5 & -3 & 0 & 0 \\ -8 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is a generalized inverse of **A**. The solution of the form (12) is

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z} \\ &= \begin{bmatrix} 6 \\ -8 \\ 0 \\ 0 \end{bmatrix} + \left\{ \begin{bmatrix} 1 & 0 & -1 & -29 \\ 0 & 1 & 2 & 47 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} - \mathbf{I} \right\} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \\ &= \begin{bmatrix} 6 - z_3 - 29z_4 \\ -8 + 2z_3 + 47z_4 \\ -z_3 \\ -z_4 \end{bmatrix} \end{aligned} \tag{14}$$

where $z_3$ and $z_4$ are arbitrary. This means that (13) is a solution to (12) no matter what the given values of $z_3$ and $z_4$ are. For example putting $z_3 = z_4 = 0$ gives

$$\tilde{\mathbf{x}}_1' = \begin{bmatrix} 6 & -8 & 0 & 0 \end{bmatrix} \tag{15}$$

Setting $z_3 = -1$ and $z_4 = 2$ gives

$$\tilde{\mathbf{x}}_2' = \begin{bmatrix} -51 & 84 & 1 & -2 \end{bmatrix}. \tag{16}$$

Both of the results in (15) and (16) can be shown to satisfy (13) by direct substitution.

This is also true of the result in (14) for all $z_3$ and $z_4$.

Again, using the algorithm in Section 1b, this time using the $2 \times 2$ minor in the second and third row and column, we obtain the generalized inverse

$$\dot{\mathbf{G}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -5 & 2 & 0 \\ 0 & 13 & -5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then (12) becomes

$$\begin{aligned} \mathbf{x} &= \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z} \\ &= \begin{bmatrix} 0 \\ 4 \\ -6 \\ 0 \end{bmatrix} + \left\{ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & -11 \\ -1 & 0 & 1 & 29 \\ 0 & 0 & 0 & 0 \end{bmatrix} - \mathbf{I} \right\} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \\ &= \begin{bmatrix} -\dot{z}_1 \\ 4 + 2\dot{z}_1 - 11\dot{z}_4 \\ -6 - \dot{z}_1 + 29\dot{z}_4 \\ -\dot{z}_4 \end{bmatrix} \end{aligned} \tag{17}$$

for arbitrary values $\dot{z}_1$ and $\dot{z}_4$. The reader may show that this too satisfies (13). □

### c. Properties of Solutions

One might ask about the relationship, if any, between the two solutions (14) and (17) found by using the two generalized inverses $\mathbf{G}$ and $\dot{\mathbf{G}}$. Both satisfy (13) for an infinite number of sets of values of $z_3$, $z_4$ and $\dot{z}_1, \dot{z}_4$. The basic question is: do the two solutions generate, though allocating different sets of values to the arbitrary values $z_3$ and $z_4$ in $\tilde{\mathbf{x}}$ and $\dot{z}_1$ and $\dot{z}_4$ in $\dot{\mathbf{x}}$, the same series of vectors satisfying $\mathbf{A}\mathbf{x} = \mathbf{y}$? The answer is "yes" because on substituting $\dot{z}_1 = -6 + z_3 + 29z_4$ and $\dot{z}_4 = z_4$ into (17) yields the solution in (14). Hence, (14) and (17) generate the same sets of solutions.

Likewise, the relationship between solutions using $\mathbf{G}$ and those using $\dot{\mathbf{G}}$ is that on substituting $\mathbf{z} = (\mathbf{G} - \dot{\mathbf{G}})\mathbf{y} + (\mathbf{I} - \dot{\mathbf{G}}\mathbf{A})\dot{\mathbf{z}}$ into (12) and noting by Theorem 1 that $\mathbf{G}\mathbf{A}\mathbf{G}\mathbf{y} = \mathbf{G}\mathbf{A}\dot{\mathbf{G}}\mathbf{y}$ $\tilde{\mathbf{x}}$ reduces to $\dot{\mathbf{x}}$.

A stronger result which concerns generation of all solutions from $\tilde{\mathbf{x}}$ is contained in the following theorem.

**Theorem 3** For the consistent equations $\mathbf{A}\mathbf{x} = \mathbf{y}$, all solutions are, for any specific $\mathbf{G}$ generated by $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z}$ for arbitrary $\mathbf{z}$.

*Proof.* Let $\mathbf{x}^*$ be any solution to $\mathbf{A}\mathbf{x} = \mathbf{y}$. Choose $\mathbf{z} = (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{x}^*$. Then

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z} = \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})(\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{x}^* \\ &= \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A}\mathbf{G}\mathbf{A} - \mathbf{G}\mathbf{A} - \mathbf{G}\mathbf{A} + \mathbf{I})\mathbf{x}^* \\ &= \mathbf{G}\mathbf{y} + (\mathbf{I} - \mathbf{G}\mathbf{A})\mathbf{x}^* = \mathbf{G}\mathbf{y} + \mathbf{x}^* - \mathbf{G}\mathbf{A}\mathbf{x}^* \\ &= \mathbf{G}\mathbf{y} + \mathbf{x}^* - \mathbf{G}\mathbf{y} = \mathbf{x}^* \end{aligned}$$

applying Theorem 1. ■

The importance of this theorem is that we need to derive only one generalized inverse of $\mathbf{A}$ to be able to generate all solutions to $\mathbf{Ax} = \mathbf{y}$. There are no solutions other than those that can be generated from $\tilde{\mathbf{x}}$.

Having established a method for solving linear equations and showing that they can have an infinite number of solutions, we ask two questions:

(i) What relationships exist among the solutions?

(ii) To what extent are the solutions linearly independent (LIN)? (A discussion of linear independence and dependence is available in Section 5 of Gruber (2014) or any standard matrix or linear algebra textbook.)

Since each solution is a vector of order $q$, there can of course be no more than $q$ LIN solutions. In fact, there are fewer, as Theorem 4 shows.

**Theorem 4** When $\mathbf{A}$ is a matrix of $q$ columns and rank $r$, and when $\mathbf{y}$ is a non-null vector, the number of LIN solutions to the consistent equations $\mathbf{Ax} = \mathbf{y}$ is $q - r + 1$.

To establish this theorem we need the following Lemma.

**Lemma 1** Let $\mathbf{H} = \mathbf{GA}$ where the rank of $\mathbf{A}$, denoted by $r(\mathbf{A})$ is $r$, that is, $r(\mathbf{A}) = r$; and $\mathbf{A}$ has $q$ columns. Then $\mathbf{H}$ is idempotent (meaning that $\mathbf{H}^2 = \mathbf{H}$) with rank $r$ and

$$r(\mathbf{I} - \mathbf{H}) = q - r.$$

*Proof.* To show that $\mathbf{H}$ is idempotent, notice that $\mathbf{H}^2 = \mathbf{GAGA} = \mathbf{GA} = \mathbf{H}$. Furthermore, by the rule for the rank of a product matrix (See Section 6 of Gruber (2014)), $r(\mathbf{H}) = r(\mathbf{GA}) \leq r(\mathbf{A})$. Similarly, because $\mathbf{AH} = \mathbf{AGA} = \mathbf{A}$, $r(\mathbf{H}) \geq r(\mathbf{AH}) = r(\mathbf{A})$. Therefore, $r(\mathbf{H}) = r(\mathbf{A}) = r$. Since $\mathbf{H}$ is idempotent, we have that $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}$. Thus, $\mathbf{I} - \mathbf{H}$ is also idempotent of order $q$. The eigenvalues of an idempotent matrix can be shown to be zero or one. The rank of a matrix corresponds to the number of non-zero eigenvalues. The trace of an idempotent matrix is the number of non-zero eigenvalues. Thus, $r(\mathbf{I} - \mathbf{H}) = \text{tr}(\mathbf{I} - \mathbf{H}) = q - \text{tr}(\mathbf{H}) = q - r$. ■

*Proof of Theorem 4.* Writing $\mathbf{H} = \mathbf{GA}$, the solutions to $\mathbf{Ax} = \mathbf{y}$ are from Theorem 2

$$\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{GA} - \mathbf{I})\mathbf{z}.$$

From Lemma 1, $r(\mathbf{I} - \mathbf{H}) = q - r$. As a result, there are only $(q - r)$ arbitrary elements in $(\mathbf{H} - \mathbf{I})\mathbf{z}$. The other $r$ elements are linear combinations of those $q - r$. Therefore, there only $(q - r)$ LIN vectors $(\mathbf{H} - \mathbf{I})\mathbf{z}$ and using them in $\tilde{\mathbf{x}}$ gives $(q - r)$ LIN solutions. For $i = 1, 2, \ldots, q - r$ let $\tilde{\mathbf{x}}_i = \mathbf{Gy} + (\mathbf{H} - \mathbf{I})\mathbf{z}_i$ be these solutions. Another solution is $\tilde{\mathbf{x}} = \mathbf{Gy}$.

Assume that this solution is linearly dependent on the $\tilde{\mathbf{x}}_i$. Then, for scalars $\lambda_i$, $i = 1, 2, \ldots, q - r$, not all of which are zero,

$$\mathbf{Gy} = \sum_{i=1}^{q-r} \lambda_i \tilde{\mathbf{x}}_i = \sum_{i=1}^{q-r} \lambda_i [\mathbf{Gy} + (\mathbf{H} - \mathbf{I})\mathbf{z}_i]$$
$$= \mathbf{Gy} \sum_{i=1}^{q-r} \lambda_i + \sum_{i=1}^{q-r} \lambda_i [(\mathbf{H} - \mathbf{I})\mathbf{z}_i]. \tag{18}$$

The left-hand side of (18) contains no $\mathbf{z}$'s. Therefore, for the last expression on the right-hand side of (18), the second term is zero. However, since the $(\mathbf{H} - \mathbf{I})\mathbf{z}_i$ are LIN, this can be true only if each of the $\lambda_i$ is zero. This means that (18) is no longer true for some $\lambda_i$ non-zero. Therefore, $\mathbf{Gy}$ is independent of the $\tilde{\mathbf{x}}_i$ so that $\mathbf{Gy}$ and $\tilde{\mathbf{x}}_i$ for $i = 1, 2, \ldots, q - r$ form a set of $(q - r + 1)$ LIN solutions. When $q = r$, there is but one solution corresponding to the existence of $\mathbf{A}^{-1}$, and that one solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. ■

Theorem 4 means that $\tilde{\mathbf{x}} = \mathbf{Gy}$ and $\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{H} - \mathbf{I})\mathbf{z}$ for $(q - r)$ LIN vectors $\mathbf{z}$ are LIN solutions to $\mathbf{Ax} = \mathbf{y}$. All other solutions will be linear combinations of these $(q - r + 1)$ solutions. Theorem 5 presents a way of constructing solutions in terms of other solutions.

**Theorem 5** If $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_s$ are any $s$ solutions of the consistent equations $\mathbf{Ax} = \mathbf{y}$ for which $\mathbf{y} \neq \mathbf{0}$, then any linear combination of these equations $\mathbf{x}^* = \sum_{i=1}^{s} \lambda_i \tilde{\mathbf{x}}_i$ is also a solution of the equations if and only if $\sum_{i=1}^{s} \lambda_i = 1$.

*Proof.* Since

$$\mathbf{x}^* = \sum\nolimits_{i=1}^{s} \lambda_i \tilde{\mathbf{x}}_i,$$

it follows that

$$\mathbf{Ax}^* = \mathbf{A} \sum_{i=1}^{s} \lambda_i \tilde{\mathbf{x}}_i = \sum_{i=1}^{s} \lambda_i \mathbf{A} \tilde{\mathbf{x}}_i.$$

Since $\tilde{\mathbf{x}}_i$ is a solution, for all $i$, $\mathbf{A}\tilde{\mathbf{x}}_i = \mathbf{y}$. This yields

$$\mathbf{Ax}^* = \sum_{i=1}^{s} \lambda_i \mathbf{y} = \mathbf{y} \left( \sum_{i=1}^{s} \lambda_i \right). \tag{19}$$

Now if $\mathbf{x}^*$ is a solution of $\mathbf{Ax} = \mathbf{y}$, then $\mathbf{Ax}^* = \mathbf{y}$ and by comparison with (19), this means, $\mathbf{y}$ being non-null, that $\sum_{i=1}^{s} \lambda_i = 1$. Conversely, if $\sum_{i=1}^{s} \lambda_i = 1$, equation (19) implies that $\mathbf{Ax}^* = \mathbf{y}$, meaning that $\mathbf{x}^*$ is a solution. This establishes the theorem. ■

Notice that Theorem 5 is in terms of any $s$ solutions. Hence, for any number of solutions whether LIN or not, any linear combination of them is itself a solution provided that the coefficients in that combination sum to unity.

**Corollary 5.1** When $\mathbf{y} = \mathbf{0}$, $\mathbf{Gy} = 0$ and there are only $q - r$ LIN solutions to $\mathbf{Ax} = \mathbf{0}$. Furthermore, for any values of the $\lambda_i$'s, $\sum_{i=1}^{s} \lambda_i \tilde{\mathbf{x}}_i$ is a solution to $\mathbf{Ax} = \mathbf{0}$.

**Example 7** Continuation of Example 6

For $\mathbf{A}$ defined in Example 6, the rank $r = 2$. Therefore, there are $q - r + 1 = 4 - 2 + 1 = 3$ LIN solutions to (13). Two are shown in (14) and (15) with (14) being the solution $\mathbf{Gy}$ when the value $\mathbf{z} = \mathbf{0}$ is used. Another solution putting $\mathbf{z}' = \begin{bmatrix} 0 & 0 & -1 & 0 \end{bmatrix}$ into (14) is

$$\tilde{\mathbf{x}}_3' = \begin{bmatrix} -23 & 39 & 0 & -1 \end{bmatrix}.$$

Thus, $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2,$ and $\tilde{\mathbf{x}}_3$ are LIN solutions and any other solution will be a combination of these three. For example, with $\mathbf{z}' = \begin{bmatrix} -23 & 39 & 0 & -1 \end{bmatrix}$, the solution (14) becomes

$$\tilde{\mathbf{x}}_4 = \begin{bmatrix} 7 & -10 & 1 & 0 \end{bmatrix}.$$

It can be seen that

$$\tilde{\mathbf{x}}_4 = 2\tilde{\mathbf{x}}_1 + \tilde{\mathbf{x}}_2 - 2\tilde{\mathbf{x}}_3.$$

The coefficients on the right-hand side of the above linear combination sums to unity in accordance with Theorem 5.

A final theorem is related to an invariance property of the elements of a solution. It is important to the study of linear models because of its relationship with the concept of estimability discussed in Chapter 5. Without worrying about the details of estimability here, we give the theorem and refer to it later as needed. The theorem is due to C. R. Rao (1962). It concerns linear combinations of the elements of a solution vector. Certain combinations are invariant to whatever solution is used.

**Theorem 6** The value of $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to whatever solution is of $\mathbf{Ax} = \mathbf{y}$ is used for $\tilde{\mathbf{x}}$ if and only if $\mathbf{k}'\mathbf{H} = \mathbf{k}'$(where $\mathbf{H} = \mathbf{GA}$ and $\mathbf{AGA} = \mathbf{A}$).

*Proof.* For a solution $\tilde{\mathbf{x}}$ given by Theorem 2

$$\mathbf{k}'\tilde{\mathbf{x}} = \mathbf{k}'\mathbf{Gy} + \mathbf{k}'(\mathbf{H} - \mathbf{I})\mathbf{z}.$$

This is independent of the arbitrary $\mathbf{z}$ if $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. Since any solution can be put in the form $\tilde{\mathbf{x}}$ by the appropriate choice of $\mathbf{z}$, the value of $\mathbf{k}'\tilde{\mathbf{x}}$ for any $\tilde{\mathbf{x}}$ is $\mathbf{k}'\mathbf{G}\mathbf{y}$ provided that $\mathbf{k}'\mathbf{H} = \mathbf{k}'$.

It may not be entirely clear that when $\mathbf{k}'\mathbf{H} = \mathbf{k}'$, the value of $\mathbf{k}'\tilde{\mathbf{x}} = \mathbf{k}'\mathbf{G}\mathbf{y}$ is invariant to the choice of $\mathbf{G}$. We therefore clarify this point. First, by Theorem 4, there are $(q - r + 1)$ LIN solutions of the form $\tilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z}$. Let these solutions be $\tilde{\mathbf{x}}_i$ for $i = 1, 2, \ldots, q - r + 1$. Suppose that for some other generalized inverse, $\mathbf{G}^*$ we have a solution

$$\mathbf{x}^* = \mathbf{G}^*\mathbf{y} + (\mathbf{H}^* - \mathbf{I})\mathbf{z}^*.$$

Then, since the $\tilde{\mathbf{x}}_i$ are a LIN set of $(q - r + 1)$ solutions $\mathbf{x}^*$ must be a linear combination of them. This means that there is a set of scalars $\lambda_i$ for $i = 1, 2, \ldots, q - r + 1$ such that

$$\mathbf{x}^* = \sum_{i=1}^{q-r+1} \lambda_i \tilde{\mathbf{x}}_i$$

where not all of the $\lambda_i$ s are zero. Furthermore, by Theorem 5, $\sum_{i=1}^{q-r+1} \lambda_i = 1$.

Proving the sufficiency part of the theorem demands showing that $\mathbf{k}'\tilde{\mathbf{x}}$ is the same for all solutions $\tilde{\mathbf{x}}$ when $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. Note that when $\mathbf{k}'\mathbf{H} = \mathbf{k}'$,

$$\mathbf{k}'\tilde{\mathbf{x}} = \mathbf{k}'\mathbf{H}\tilde{\mathbf{x}} = \mathbf{k}'\mathbf{H}\mathbf{G}\mathbf{y} + \mathbf{k}'(\mathbf{H}^2 - \mathbf{H})\mathbf{z} = \mathbf{k}'\mathbf{H}\mathbf{G}\mathbf{y} = \mathbf{k}'\mathbf{G}\mathbf{y}.$$

Therefore, $\mathbf{k}'\tilde{\mathbf{x}}_i = \mathbf{k}'\mathbf{G}\mathbf{y}$ for all $i$, and

$$\begin{aligned}\mathbf{k}'\mathbf{x}^* &= \mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i \tilde{\mathbf{x}}_i = \sum_{i=1}^{q-r+1} \lambda_i \mathbf{k}\tilde{\mathbf{x}}_i = \sum_{i=1}^{q-r+1} \lambda_i \mathbf{k}\mathbf{G}\mathbf{y} = \mathbf{k}'\mathbf{G}\mathbf{y}\left(\sum_{i=1}^{q-r+1} \lambda_i\right) \\ &= \mathbf{k}'\mathbf{G}\mathbf{y} = \mathbf{k}'\tilde{\mathbf{x}}_i.\end{aligned}$$

That means that for any solution at all $\mathbf{k}'\tilde{\mathbf{x}} = \mathbf{k}'\mathbf{G}\mathbf{y}$ if $\mathbf{k}'\mathbf{H} = \mathbf{k}'$.

To prove the necessity part of the theorem, choose $\mathbf{z}^* = \mathbf{0}$ in $\mathbf{x}^*$. Then

$$\begin{aligned}\mathbf{k}'\mathbf{x}^* = \mathbf{k}'\mathbf{G}\mathbf{y} = \mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i \tilde{\mathbf{x}}_i &= \mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i [\mathbf{G}\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z}_i] \\ &= \mathbf{k}'\mathbf{G}\mathbf{y}\left(\sum_{i=1}^{q-r+1} \lambda_i\right) + \mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i (\mathbf{H} - \mathbf{I})\mathbf{z}_i \\ &= \mathbf{k}'\mathbf{G}\mathbf{y} + \mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i (\mathbf{H} - \mathbf{I})\mathbf{z}_i.\end{aligned}$$

Hence, $\mathbf{k}' \sum_{i=1}^{q-r+1} \lambda_i(\mathbf{H} - \mathbf{I})\mathbf{z}_i = 0$. However, the $\lambda_i$ are not all zero and the $(\mathbf{H}-\mathbf{I})\mathbf{z}_i$ are LIN. Therefore, this last equation can be true only if $\mathbf{k}'(\mathbf{H} - \mathbf{I}) = 0$, that is, $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. Hence, for any solution $\mathbf{x}^*$, $\mathbf{k}'\mathbf{x}^* = \mathbf{k}'\mathbf{G}\mathbf{y}$ if and only if $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. This proves the theorem conclusively. ■

**Example 8** Illustration of the Invariance Principle

In deriving (14) in Example 6, we have that

$$\mathbf{H} = \mathbf{G}\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 & -29 \\ 0 & 1 & 2 & 47 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and for

$$\mathbf{k}' = \begin{bmatrix} 3 & 2 & 1 & 7 \end{bmatrix}, \tag{20}$$

it will be found that $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. Therefore, $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to whatever solution is used for $\tilde{\mathbf{x}}$. Thus, from (15) and (16)

$$\mathbf{k}'\tilde{\mathbf{x}}_1 = 3(6) + 2(-8) + 1(0) + 7(0) = 2$$

and

$$\mathbf{k}'\tilde{\mathbf{x}}_2 = 3(-51) + 2(84) + 1(1) + 7(-2) = 2.$$

In general, from (14)

$$\mathbf{k}'\tilde{\mathbf{x}} = 3(6 - z_3 - 29z_4) + 2(-8 + 2z_3 + 47z_4) + 1(-z_3) + 7(-z_4) = 2.$$

Likewise, $\mathbf{k}'\dot{\mathbf{x}}$ has the same value. From (17)

$$\mathbf{k}'\dot{\mathbf{x}} = 3(-\dot{z}_1) + 2(4 + 2\dot{z}_1 - 11\dot{z}_4) + 1(-6 - \dot{z}_1 + 29\dot{z}_4) + 7(-\dot{z}_4) = 2.$$

There are of course many values of $\mathbf{k}'$ that satisfy $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. For each of these $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to the choice of $\tilde{\mathbf{x}}$. For two such vectors $\mathbf{k}_1'$ and $\mathbf{k}_2'$ say $\mathbf{k}_1'\tilde{\mathbf{x}}$ and $\mathbf{k}_2'\tilde{\mathbf{x}}$ are different but each has a value that is the same for all values of $\tilde{\mathbf{x}}$. Thus, in the example $\mathbf{k}_1'\mathbf{H} = \mathbf{k}_1'$, where

$$\mathbf{k}_1' = \begin{bmatrix} 1 & 2 & 3 & 65 \end{bmatrix}$$

is different from (20) and

$$\mathbf{k}_1'\tilde{\mathbf{x}}_1 = 1(6) + 2(-8) + 3(0) + 65(0) = -10$$

is different from $\mathbf{k}'\tilde{\mathbf{x}}$ for $\mathbf{k}'$ of (20). However, for every $\tilde{\mathbf{x}}$, $\mathbf{k}_1'\tilde{\mathbf{x}}_1 = -10$. □

It was shown in Theorem 6 that the invariance of $\mathbf{k}'\tilde{\mathbf{x}}$ to $\tilde{\mathbf{x}}$ holds for any $\mathbf{k}'$ provided that $\mathbf{k}'\mathbf{H} = \mathbf{k}'$. Two corollaries of the theorem follow.

**Corollary 6.1** The linear combination $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to $\tilde{\mathbf{x}}$ for $\mathbf{k}'$ of the form $\mathbf{k}' = \mathbf{w}'\mathbf{H}$ for arbitrary $\mathbf{w}'$.

*Proof.* We have that $\mathbf{k}'\mathbf{H} = \mathbf{w}'\mathbf{H}^2 = \mathbf{w}'\mathbf{GAGA} = \mathbf{w}'\mathbf{GA} = \mathbf{w}'\mathbf{H} = \mathbf{k}'$. ■

**Corollary 6.2** There are only $r$ LIN vectors $\mathbf{k}'$ for which $\mathbf{k}'\tilde{\mathbf{x}}$ is invariant to $\tilde{\mathbf{x}}$.

*Proof.* Since $r(\mathbf{H}) = r$, there are in $\mathbf{k}' = \mathbf{w}'\mathbf{H}$ of order $q$ exactly $q - r$ elements that are linear combinations of the other $r$. Therefore, for arbitrary vectors $\mathbf{w}'$ there are only $r$ LIN vectors $\mathbf{k}' = \mathbf{w}'\mathbf{H}$. ■

We will return to this point in Chapter 5 when we discuss estimable functions.

The concept of generalized inverse has now been defined and its use in solving linear equations explained. Next, we briefly discuss the generalized inverse itself, its various definitions and some of its properties. Extensive review of generalized inverses and their applications is to be found in Boullion and Odell (1968) and the approximately 350 references there. A more recent reference on generalized inverses is Ben-Israel and Greville (2003).

## 3. THE PENROSE INVERSE

Penrose (1955) in extending the work of Moore (1920), shows that for every matrix $\mathbf{A}$, there is a unique matrix $\mathbf{K}$ which satisfies the following conditions:

$$\begin{aligned} &\mathbf{AKA} = \mathbf{A} && \text{(i)} \\ &\mathbf{KAK} = \mathbf{K} && \text{(ii)} \\ &(\mathbf{KA})' = \mathbf{KA} && \text{(iii)} \\ &(\mathbf{AK})' = \mathbf{AK} && \text{(iv)} \end{aligned} \tag{21}$$

Such generalized inverses $\mathbf{K}$ will be referred to as Moore–Penrose inverses. We will show how to find them and prove that every matrix has a unique Moore–Penrose inverse.

Condition (i) states that $\mathbf{K}$ is a generalized inverse of $\mathbf{A}$. Condition (ii) states that $\mathbf{A}$ is a generalized inverse of $\mathbf{K}$**.** In Section 4, we will give an example to show that in general, Condition (i) does not imply condition (ii). Conditions (iii) and (iv) state that $\mathbf{KA}$ and $\mathbf{AK}$**,** respectively, are symmetric matrices. There are generalized inverses that satisfy one or more of conditions (ii), (iii), and (iv) but not all of them. We will give examples of such generalized inverses in Section 4.

In Section 2, we showed how to obtain a generalized inverse using the singular value decomposition. These generalized inverses satisfy all four of the above conditions and, as a result, are Moore–Penrose inverses. Although a matrix has infinitely many generalized inverses it has only one Moore–Penrose inverse. We show this below.

**Theorem 7** Let $\mathbf{A}$ be a matrix with singular value decomposition $\mathbf{S}'\Lambda^{1/2}\mathbf{U}'$. Then the generalized inverse $\mathbf{K} = \mathbf{U}\Lambda^{-1/2}\mathbf{S}$ is the Moore–Penrose inverse.

*Proof.* We have already shown that $\mathbf{K}$ is in fact a generalized inverse. To establish the second Penrose condition we have

$$\mathbf{KAK} = \mathbf{U}\Lambda^{-1/2}\mathbf{SS}'\Lambda^{1/2}\mathbf{U}'\mathbf{U}\Lambda^{-1/2}\mathbf{S} = \mathbf{U}\Lambda^{-1/2}\mathbf{S} = \mathbf{K}.$$

Now

$$\mathbf{KA} = \mathbf{U}\Lambda^{-1/2}\mathbf{SS}'\Lambda^{1/2}\mathbf{U}' = \mathbf{UU}'.$$

and

$$\mathbf{AK} = \mathbf{S}'\Lambda^{1/2}\mathbf{U}'\mathbf{U}\Lambda^{-1/2}\mathbf{S} = \mathbf{S}'\mathbf{S}.$$

Since $\mathbf{UU}'$ and $\mathbf{SS}'$ are symmetric matrices, conditions (iii) and (iv) are established. ■

The generalized inverse in Example 5 is the Moore–Penrose inverse of $\mathbf{A}$ there.

We have thus established the existence of the Moore–Penrose inverse. We now show that it is unique.

**Theorem 8** The Moore–Penrose inverse is unique.

*Proof.* The proof consists of showing that for a given matrix, there can be only one matrix that satisfies the four conditions. First, from condition (i) and (iii)

$\mathbf{A} = \mathbf{AKA} = \mathbf{A}(\mathbf{KA})' = \mathbf{AA}'\mathbf{K}'$ and by transposition

$$\mathbf{KAA}' = \mathbf{A}'. \tag{22}$$

Also if $\mathbf{AA}'\mathbf{K}' = \mathbf{A}$, then $\mathbf{KAA}'\mathbf{K}' = \mathbf{KA}(\mathbf{KA})' = \mathbf{KA}$ so that $\mathbf{KA}$ is symmetric.

Also $\mathbf{AKA} = \mathbf{A}(\mathbf{KA})' = \mathbf{AA}'\mathbf{K}' = \mathbf{A}$. Thus (22) is equivalent to (i) and (iii).

Likewise, from condition (ii) and (iv), we can show in a similar way that an equivalent identity is

$$\mathbf{KK}'\mathbf{A}' = \mathbf{K}. \tag{23}$$

Suppose that $\mathbf{K}$ is not unique. Assume some other matrix $\mathbf{M}$ satisfies the Penrose conditions. From conditions (i) and (iv), we have

$$\mathbf{AA'M} = \mathbf{A'} \tag{24}$$

and from conditions (ii) and (iii)

$$\mathbf{A'M'M} = \mathbf{M}. \tag{25}$$

We then have that using (22)–(25),

$$\mathbf{K} = \mathbf{KK'A'} = \mathbf{KK'A'AM} = \mathbf{KAM} = \mathbf{KAA'M'M} = \mathbf{A'MM'} = \mathbf{M}.$$

This establishes uniqueness. ■

We now give another method for finding the Moore–Penrose inverse based on the Cayley–Hamilton Theorem (see, for example, Searle (1966), C. R. Rao (1973), and Gruber (2014)). The Cayley–Hamilton theorem states that a square matrix satisfies its characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. To show this, we need two lemmas.

**Lemma 2** If the matrix $\mathbf{X'X} = 0$, then $\mathbf{X} = 0$.

*Proof.* If the matrix $\mathbf{X'X} = 0$, then the sums of squares of the elements of each row are zero so that the elements themselves are zero. ■

**Lemma 3** The identity $\mathbf{PX'X} = \mathbf{QX'X}$ implies that $\mathbf{PX'} = \mathbf{QX'}$.

*Proof.* Apply Lemma 2 to

$$(\mathbf{PX'X} - \mathbf{QX'X})(\mathbf{P} - \mathbf{Q}) = (\mathbf{PX'} - \mathbf{QX'})(\mathbf{PX'} - \mathbf{QX'})' = 0.$$

We will give an alternative proof that uses the singular value decomposition of $\mathbf{X} = \mathbf{S'}\boldsymbol{\Lambda}^{1/2}\mathbf{U'}$. We have that $\mathbf{PX'X} = \mathbf{QX'X}$ implies that $\mathbf{PU}\Lambda^{1/2}\mathbf{SS'}\Lambda^{1/2}\mathbf{U'} = \mathbf{QU}\Lambda^{1/2}\mathbf{SS'}\Lambda^{1/2}\mathbf{U'}$.

Multiply both sides of this equation by $\mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{S}$ and obtain $\mathbf{PU}\Lambda^{1/2}\mathbf{SS'S} = \mathbf{QU}\Lambda^{1/2}\mathbf{SS'S}$. Since $\mathbf{SS'} = \mathbf{I}$, we have

$$\mathbf{PU}\Lambda^{1/2}\mathbf{S} = \mathbf{QU}\Lambda^{1/2}\mathbf{S} \text{ or } \mathbf{PX'} = \mathbf{QX'}.$$

■

We now assume that

$$\mathbf{K} = \mathbf{TA'} \tag{26}$$

for some matrix $\mathbf{T}$. Then (22) is satisfied if $\mathbf{T}$ satisfies

$$\mathbf{TA'AA'} = \mathbf{A'}; \tag{27}$$

and since satisfaction of (22) implies satisfaction of conditions (i) and (iii). Thus, $\mathbf{AKA} = \mathbf{A}$ and $\mathbf{A'K'A'} = \mathbf{A'}$. As a result, $\mathbf{TA'K'A'} = \mathbf{TA'}$, or from (26), we get (23).

However, (23) is equivalent to Penrose conditions (ii) and (iv) so $\mathbf{K}$ as defined in (26) for $\mathbf{T}$ that satisfies (27).

We now derive a suitable **T.** Notice that the matrix $\mathbf{A'A}$ and all of its powers are square. By the Cayley–Hamilton Theorem, for some integer $t$, there exists a series of scalars $\lambda_1, \lambda_2, \ldots, \lambda_t$ not all zero, such that

$$\lambda_1\mathbf{A'A} + \lambda_2(\mathbf{A'A})^2 + \cdots + \lambda_t(\mathbf{A'A})^t = 0.$$

If $\lambda_r$ is the first $\lambda$ in this identity that is non-zero then $\mathbf{T}$ is defined as

$$\mathbf{T} = (-1/\lambda_r)[\lambda_{r+1}\mathbf{I} + \lambda_{r+2}(\mathbf{A'A}) + \cdots + \lambda_t(\mathbf{A'A})^{t-r-1}]. \tag{28}$$

To show that this satisfies (27) note that by direct multiplication

$$\begin{aligned}\mathbf{T}(\mathbf{A'A})^{r+1} &= (-1/\lambda_r)[\lambda_{r+1}(\mathbf{A'A})^{r+1} + \lambda_{r+2}(\mathbf{A'A})^{r+2} + \cdots \lambda_t(\mathbf{A'A})^t] \\ &= (-1/\lambda_r)[-\lambda_1\mathbf{A'A} - \lambda_2(\mathbf{A'A})^2 - \cdots \lambda_r(\mathbf{A'A})^r].\end{aligned}$$

Since by definition $\lambda_r$ is the first non-zero $\lambda$ in the series $\lambda_1, \lambda_2, \ldots$, the above reduces to

$$\mathbf{T}(\mathbf{A'A})^{r+1} = (\mathbf{A'A})^r. \tag{29}$$

Repeated use of Lemma 3 reduces this to (27). Thus, $\mathbf{K} = \mathbf{TA'}$ with $\mathbf{T}$ as defined in (28) satisfies (27) and hence is the unique generalized inverse satisfying all four of the Penrose conditions.

**Example 9** Finding a Moore–Penrose Inverse using the Cayley–Hamilton Theorem
For

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 1 \\ -1 & 0 & -2 \\ 1 & 2 & 0 \end{bmatrix}, \text{ we have } \mathbf{A'A} = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 5 & -1 \\ 4 & -1 & 9 \end{bmatrix}.$$

Finding the characteristic equation $66\lambda - 17\lambda^2 + \lambda^3 = 0$ and employing the Cayley–Hamilton theorem, we have

$$66(\mathbf{A'A}) - 17(\mathbf{A'A})^2 + (\mathbf{A'A})^3 = 0.$$

Then,

$$\mathbf{T} = (-1/66)(-17\mathbf{I} + \mathbf{A}'\mathbf{A}) = (1/66)\begin{bmatrix} 14 & -2 & -4 \\ -2 & 12 & 1 \\ -4 & 1 & 8 \end{bmatrix}$$

and $\mathbf{K} = \mathbf{TA}' = (1/66)\begin{bmatrix} 6 & -2 & -6 & 10 \\ 0 & -11 & 0 & 22 \\ 12 & 7 & -12 & -2 \end{bmatrix}$ is the Penrose inverse of $\mathbf{A}$ satisfying 21. □

Graybill et al. (1966) suggests an alternative procedure for deriving $\mathbf{K}$. Their method is to find X and $\mathbf{Y}$ such that

$$\mathbf{AA}'\mathbf{X}' = \mathbf{A} \quad \text{and} \quad \mathbf{A}'\mathbf{AY} = \mathbf{A}' \tag{30}$$

and then

$$\mathbf{K} = \mathbf{XAY}. \tag{31}$$

Proof that $\mathbf{K}$ satisfies all four Penrose axioms depends on using (30) and Lemma 3 to show that $\mathbf{AXA} = \mathbf{A} = \mathbf{AYA}$. (See Exercise 28.)

## 4. OTHER DEFINITIONS

It is clear that the Penrose inverse $\mathbf{K}$ is not easy to compute, especially when $\mathbf{A}$ has many columns or irrational eigenvalues because either finding the singular value decomposition or using the Cayley–Hamilton theorem can be quite tedious. As has already been shown, only the first Penrose condition needs to be satisfied to have a matrix useful for solving linear equations. Furthermore, in pursuing the topic of linear models, this is the only condition that is really needed. For this reason, a generalized inverse has been defined as any matrix that satisfies $\mathbf{AGA} = \mathbf{A}$. This definition will be retained throughout the book. Nevertheless, a variety of names will be found throughout the literature, both for $\mathbf{G}$ and for other matrices satisfying fewer than all four of the Penrose conditions. There are five such possibilities as detailed in Table 1.1.

In the notation of Table 1.1 $\mathbf{A}^{(g)} = \mathbf{G}$, the generalized inverse already defined and discussed, and $\mathbf{A}^{(p)} = \mathbf{K}$, the Moore–Penrose inverse. This has also been called the pseudo inverse and the p-inverse by various authors. The Software package Mathematica computes the Moore–Penrose inverse of $\mathbf{A}$ in response to the input PseudoInverse[A]. The suggested definition of normalized generalized inverse in Table 1.1 is not universally accepted. As given there it is used by Urquhart (1968), whereas Goldman and Zelen (1964), call it a "weak" generalized inverse. An example of such a matrix is a left inverse $\mathbf{L}$ such that $\mathbf{LA} = \mathbf{I}$. Rohde (1966) has also used the description "normalized" (we use reflexive least square) for a matrix satisfying

**TABLE 1.1 Suggested Names for Matrices Satisfying Some or All of the Penrose Conditions**

| Conditions Satisfied (Eq. 21) | Name of Matrix | Symbol |
|---|---|---|
| i | Generalized inverse | $\mathbf{A}^{(g)}$ |
| i and ii | Reflexive generalized inverse | $\mathbf{A}^{(r)}$ |
| i and iii | Mininum norm generalized inverse | $\mathbf{A}^{(mn)}$ |
| i and iv | Least-square generalized inverse | $\mathbf{A}^{(ls)}$ |
| i, ii, and iii | Normalized generalized inverse | $\mathbf{A}^{(n)}$ |
| i, ii, and iv | Reflexive least square Generalized inverse | $\mathbf{A}^{(rls)}$ |
| i, ii, iii, and iv | Moore–Penrose inverse | $\mathbf{A}^{(p)}$ |

conditions (i), (ii), and (iv). An example of this kind of matrix is a right inverse $\mathbf{R}$ for which $\mathbf{AR} = \mathbf{I}$.

The generalized inverses obtained in Section 1 by diagonalization or the algorithm are reflexive. See Exercise 27.

Let $\mathbf{x} = \mathbf{Gy}$ be a solution to $\mathbf{Ax} = \mathbf{y}$. The minimum norm generalized inverse is such that $\min_{\mathbf{Ax}=\mathbf{y}} \|\mathbf{x}\| = \|\mathbf{Gy}\|$. Such a generalized inverse satisfies Penrose conditions (i) and (iii). The least-square generalized inverse is the one that yields the solution $\tilde{\mathbf{x}}$ such that $\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y}\| = \inf_x \|\mathbf{Ax} - \mathbf{y}\|$. It satisfies Penrose conditions (i) and (iv). Proofs of these results are available in Gruber (2014), and Rao and Mitra (1971).

The following relationships can be established between the generalized inverses.

$$\begin{aligned} \mathbf{A}^{(r)} &= \mathbf{A}^{(g)}\mathbf{A}\mathbf{A}^{(g)} \\ \mathbf{A}^{(n)} &= \mathbf{A}'(\mathbf{A}\mathbf{A}')^{(g)} \\ \mathbf{A}^{(rls)} &= (\mathbf{A}'\mathbf{A})^{(g)}\mathbf{A}' \\ \mathbf{A}^{(p)} &= \mathbf{A}^{(n)}\mathbf{A}\mathbf{A}^{(rls)} \end{aligned} \tag{32}$$

Some general conditions for generalized inverses to be reflexive, minimum norm or least square are developed in Gruber (2014).

**Example 10** Finding Different Kinds of Generalized Inverses

As in Example 9,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 1 \\ -1 & 0 & -2 \\ 1 & 2 & 0 \end{bmatrix}, \mathbf{A}'\mathbf{A} = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 5 & -1 \\ 4 & -1 & 9 \end{bmatrix}, \text{ and } \mathbf{A}\mathbf{A}' = \begin{bmatrix} 5 & 2 & -5 & 1 \\ 2 & 2 & -2 & -2 \\ -5 & -2 & 5 & -1 \\ 1 & -2 & -1 & 5 \end{bmatrix}.$$

These three matrices have rank 2. Using the algorithm in Part 1,

$$\mathbf{A}^{(g)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since this is a reflexive generalized inverse $\mathbf{A}^{(r)} = \mathbf{A}^{(g)}$. Now,

$$(\mathbf{AA}')^{(g)} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & \frac{5}{6} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{A}^{(n)} = \mathbf{A}'(\mathbf{AA}')^{(g)} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 & 0 \\ \frac{1}{3} & -\frac{5}{6} & 0 & 0 \\ \frac{1}{3} & \frac{1}{6} & 0 & 0 \end{bmatrix}.$$

Furthermore,

$$(\mathbf{A}'\mathbf{A})^{(g)} = \begin{bmatrix} \frac{5}{11} & -\frac{2}{11} & 0 \\ -\frac{2}{11} & \frac{3}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{A}^{(rls)} = (\mathbf{A}'\mathbf{A})^{(g)}\mathbf{A}' = \begin{bmatrix} \frac{5}{11} & \frac{2}{11} & -\frac{5}{11} & \frac{1}{11} \\ -\frac{2}{11} & -\frac{3}{11} & \frac{2}{11} & \frac{4}{11} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then

$$\mathbf{A}^{(p)} = \mathbf{A}^{(n)}\mathbf{AA}^{(rls)} = (1/66)\begin{bmatrix} 6 & -2 & -6 & 10 \\ 0 & -11 & 0 & 22 \\ 12 & 7 & -12 & -2 \end{bmatrix}.$$

□

## 5. SYMMETRIC MATRICES

The study of linear models frequently leads to equations of the form $\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}$ that have to be solved for $\hat{\mathbf{b}}$. Special attention is given therefore to the properties of a generalized inverse of the symmetric matrix $\mathbf{X}'\mathbf{X}$.

### a. Properties of a Generalized Inverse

The facts summarized in Theorem 9 below will be useful. We will denote the Moore–Penrose inverse by $(\mathbf{X}'\mathbf{X})^{+}$ and any generalized inverse by $(\mathbf{X}'\mathbf{X})^{-}$

**Theorem 9** Assume that the singular value decomposition of $\mathbf{X} = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}'$. Then

(i) $\mathbf{X}'\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}'$ and $(\mathbf{X}'\mathbf{X})^{+} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}'$.

(ii) For any generalized inverse of $\mathbf{X}'\mathbf{X}$, $\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{U} = \mathbf{\Lambda}^{-1}$ and therefore $(\mathbf{X}'\mathbf{X})^{+} = \mathbf{U}\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{U}\mathbf{U}'$.

(iii) Any generalized inverse $\mathbf{G}$ of $\mathbf{X}'\mathbf{X}$ may be written in terms of the Moore–Penrose inverse as follows

$$\begin{aligned}\mathbf{G} &= (\mathbf{X}'\mathbf{X})^{+} + \mathbf{V}\mathbf{C}_1\mathbf{U}' + \mathbf{U}\mathbf{C}_1'\mathbf{V}' + \mathbf{V}\mathbf{C}_2\mathbf{V}' \\ &= \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}\begin{bmatrix}\Lambda^{-1} & \mathbf{C}_1 \\ \mathbf{C}_1' & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix},\end{aligned}$$

where $\mathbf{C}_1 = \mathbf{V}'\mathbf{G}\mathbf{U}$ and $\mathbf{C}_2 = \mathbf{V}'\mathbf{G}\mathbf{V}$

*Proof.* For (i) $\mathbf{X}'\mathbf{X} = \mathbf{U}\Lambda^{1/2}\mathbf{S}\mathbf{S}'\Lambda^{1/2}\mathbf{U}' = \mathbf{U}\Lambda\mathbf{U}'$ because $\mathbf{S}\mathbf{S}' = \mathbf{I}$. The expression $\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}'$ can be shown to satisfy the Penrose conditions.

For (ii) we have that $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$.

Then this implies that

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'. \tag{33}$$

Multiply both sides of equation (33) on the left by $\boldsymbol{\Lambda}^{-1}\mathbf{U}'$ and on the right by $\mathbf{U}\boldsymbol{\Lambda}^{-1}$. The result follows.

To establish (iii), notice that

$$\begin{aligned}\mathbf{G} &= (\mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}')\mathbf{G}(\mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}') \\ &= \mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}'\mathbf{G}\mathbf{U}\mathbf{U}' + \mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{V}\mathbf{V}' + \mathbf{V}\mathbf{V}'\mathbf{G}\mathbf{V}\mathbf{V}' \\ &= (\mathbf{X}'\mathbf{X})^{+} + \mathbf{V}\mathbf{C}_1\mathbf{U}' + \mathbf{U}\mathbf{C}_1'\mathbf{V}' + \mathbf{V}\mathbf{C}_2\mathbf{V}' \\ &= \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}^{-1} & \mathbf{C}_1 \\ \mathbf{C}_1' & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix}.\end{aligned}$$

Theorem 10 below gives some more useful properties of generalized inverses of $\mathbf{X}'\mathbf{X}$. ■

**Theorem 10** When $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$ then

(i) $\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\mathbf{X}$;

(ii) $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$; that is, $\mathbf{G}\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$;

(iii) $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to $\mathbf{G}$;

(iv) $\mathbf{X}\mathbf{G}\mathbf{X}'$ is symmetric whether G is or not.

*Proof.*

(i) By definition, $\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$. Transposition yields $\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$.

(ii) Observe that $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{S}'\boldsymbol{\Lambda}^{1/2}\mathbf{U}'\mathbf{G}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = \mathbf{S}'\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\mathbf{U}' = \mathbf{S}'\boldsymbol{\Lambda}^{1/2}\mathbf{U}' = \mathbf{X}$.

The result may also be obtained by application of Lemma 3.

(iii) Notice that $\mathbf{XGX}' = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}'\mathbf{GU}\mathbf{\Lambda}^{1/2}\mathbf{S} = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}^{1/2}\mathbf{S} = \mathbf{S}'\mathbf{S}$.

(iv) If $\mathbf{M}$ is a symmetric generalized inverse of $\mathbf{X}'\mathbf{X}$ then $\mathbf{XMX}'$ is symmetric. (For example, the Moore–Penrose inverse of $\mathbf{X}'\mathbf{X}$ is symmetric.) From (iii) $\mathbf{XGX}' = \mathbf{XMX}'$ and is thus, symmetric whether or not $\mathbf{G}$ is.

■

**Corollary 10.1** Applying part (i) of Theorem 10 to the other parts shows that

$$\mathbf{XGX}'\mathbf{X} = \mathbf{X}, \mathbf{X}'\mathbf{XGX}' = \mathbf{X}' \text{ and } \mathbf{X}'\mathbf{XG}'\mathbf{X} = \mathbf{X}'.$$

Furthermore,

$$\mathbf{XG}'\mathbf{X}' = \mathbf{XGX}' \text{ and } \mathbf{XG}'\mathbf{X}' \text{ is symmetric.}$$

It is to be emphasized that not all generalized inverses of a symmetric matrix are symmetric. This is illustrated in Example 11 below.

**Example 11** The Generalized Inverse of a Symmetric Matrix Need not be Symmetric

We can demonstrate this by applying the algorithm at the end of Section 1 to the symmetric matrix using the sub-matrix from the first two columns of the first and third rows

$$\mathbf{A}_2 = \begin{bmatrix} 2 & 2 & 6 \\ 2 & 3 & 8 \\ 6 & 8 & 22 \end{bmatrix}$$

to obtain the non-symmetric generalized inverse

$$\mathbf{G} = \begin{bmatrix} 2 & -\frac{3}{2} & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

□

Theorem 10 and Corollary 10.1 very largely enable us to avoid difficulties that this lack of symmetry of generalized inverses of $\mathbf{X}'\mathbf{X}$ might otherwise appear to involve. For example, if $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$ and $\mathbf{P}$ is some other matrix,

$$(\mathbf{PXGX}')' = \mathbf{XG}'\mathbf{X}'\mathbf{P}' = \mathbf{XGX}'\mathbf{P}'$$

not because $\mathbf{G}$ is symmetric (which in general is not) but because $\mathbf{XGX}'$ is symmetric.

□

**Example 12** Illustration of Symmetry of $\mathbf{XGX}'$

If

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ 0 & 1 & 2 \end{bmatrix},$$

then $\mathbf{X}'\mathbf{X} = \mathbf{A}_2$ from Example 11. Then

$$\mathbf{XGX}' = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & -\frac{3}{2} & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 3 & 3 & 2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

and

$$\mathbf{XG}'\mathbf{X}' = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 & -\frac{1}{2} \\ -\frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 3 & 3 & 2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

□

### b. Two More Generalized Inverses of $\mathbf{X}'\mathbf{X}$

In addition to the methods studied already, two other methods discussed by John (1964) are sometimes pertinent to linear models. They depend on the ordinary inverse of a non-singular matrix:

$$\mathbf{S}^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} = \mathbf{0} \end{bmatrix}. \tag{34}$$

The matrix $\mathbf{H}$ being used here is not the matrix $\mathbf{H} = \mathbf{GA}$ used earlier. It is being used to be consistent with John's notation. The matrix $\mathbf{X}'\mathbf{X}$ is of order $p$ and rank $p - m$. The matrix $\mathbf{H}$ is any matrix of order $m \times p$. It is of full row rank and its rows also LIN of those of $\mathbf{X}'\mathbf{X}$. In other words, the rows of $\mathbf{H}$ cannot be a linear combination of rows of $\mathbf{X}'\mathbf{X}$. (The existence of such a matrix is assured by considering $m$ vectors of order $p$ that are LIN of any set of $p - m$ LIN rows of $\mathbf{X}'\mathbf{X}$. Furthermore, if these rows constitute $\mathbf{H}$ in such a way that the $m$ LIN rows of $\mathbf{H}$ correspond in $\mathbf{S}$ to the $m$ rows of $\mathbf{X}'\mathbf{X}$ that are linear combinations of the set of $p - m$ rows then $\mathbf{S}^{-1}$ of (34) exists.) With (34) existing the two matrices

$$\mathbf{B}_{11} \text{ and } (\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1} \text{ are generalized inverses of } \mathbf{X}'\mathbf{X}. \tag{35}$$

Three useful lemmas help in establishing these results.

**Lemma 4** The matrix $\mathbf{T} = \begin{bmatrix} \mathbf{I}_r & \mathbf{U} \end{bmatrix}$ has rank $r$ for any matrix $\mathbf{U}$ of $r$ rows.

*Proof.* Elementary operations carried out on $\mathbf{T}$ to find its rank will operate on $\mathbf{I}_r$. None of these rows or columns can be made null by such operations. Therefore, $r(\mathbf{T})$ is not less than $r$. Consequently $r(\mathbf{T}) = r$. ■

**Lemma 5** If $\mathbf{X}_{N\times p}$ has rank $p - m$ for $m > 0$, then there exists a matrix $\mathbf{D}_{p\times m}$ such that $\mathbf{XD} = \mathbf{0}$ and $r(\mathbf{D}) = m$.

*Proof.* Let $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ where $\mathbf{X}_1$ is $N \times (p - m)$ of full column rank. Then the columns of $\mathbf{X}_2$ are linear combinations of the columns of $\mathbf{X}_1$ and so for some matrix $\mathbf{C}$, of order $(p - m) \times m$, the sub-matrices of $\mathbf{X}$ satisfy $\mathbf{X}_2 = \mathbf{X}_1\mathbf{C}$. Let $\mathbf{D}' = \begin{bmatrix}-\mathbf{C}' & \mathbf{I}_m\end{bmatrix}$. By Lemma 4 $\mathbf{D}'$ has rank $m$. We then have $\mathbf{XD} = \mathbf{0}$ and $r(\mathbf{D}) = m$. The Lemma is thus proved because a matrix $\mathbf{D}$ exists. ■

**Lemma 6** For $\mathbf{X}$ and $\mathbf{D}$ of Lemma 5 and $\mathbf{H}$ of order $m \times p$ with full-row rank, $\mathbf{HD}$ has full-row rank if and only if the rows of $\mathbf{H}$ are LIN of those of $\mathbf{X}$.

*Proof.* (i) Given $r(\mathbf{HD}) = m$, assume that the rows of $\mathbf{H}$ depend on those of $\mathbf{X}$ (are not LIN of $\mathbf{X}$). Then, $\mathbf{H} = \mathbf{KX}$ for some $\mathbf{K}$, and $\mathbf{HD} = \mathbf{KXD} = \mathbf{0}$. Therefore, the assumption is false and the rows of $\mathbf{H}$ are LIN of those of $\mathbf{X}$.

(ii) Given that the rows of $\mathbf{H}$ are LIN of those of $\mathbf{X}$, the matrix $\begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}$, of order $(N + m) \times p$ has full column rank. Therefore, it has a left inverse $\begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}$, say (Section 5.13 of Searle (1966)), and so $\mathbf{UX} + \mathbf{VH} = \mathbf{I}$, that is, $\mathbf{UXD} + \mathbf{VHD} = \mathbf{D}$; or $\mathbf{VHD} = \mathbf{D}$ using Lemma 5. However, $r(\mathbf{D}_{p\times m}) = m$ and $\mathbf{D}$ has a left inverse, $\mathbf{E}$, say, and so $\mathbf{EVHD} = \mathbf{I}_m$. Therefore, $r(\mathbf{HD}) \geq m$ and so because $\mathbf{HD}$ is $m \times m$, $r(\mathbf{HD}) = m$, and the lemma is proved. ■

*Proof of (35).* First it is necessary to show that in (34), $\mathbf{B}_{22} = \mathbf{0}$. From (34), we have that

$$\mathbf{X}'\mathbf{XB}_{11} + \mathbf{H}'\mathbf{B}_{21} = \mathbf{I} \quad \text{and} \quad \mathbf{X}'\mathbf{XB}_{12} + \mathbf{H}'\mathbf{B}_{22} = \mathbf{0} \tag{36}$$

$$\mathbf{HB}_{11} = \mathbf{0} \quad \text{and} \quad \mathbf{HB}_{12} = \mathbf{I}. \tag{37}$$

Pre-multiplying (36) by $\mathbf{D}'$ and using Lemmas 5 and 6 leads to

$$\mathbf{B}_{21} = (\mathbf{D}'\mathbf{H}')^{-1}\mathbf{D}' \quad \text{and} \quad \mathbf{B}_{22} = \mathbf{0}. \tag{38}$$

Then from (36) and (38),

$$\mathbf{X}'\mathbf{XB}_{11} = \mathbf{I} - \mathbf{H}'(\mathbf{D}'\mathbf{H}')^{-1}\mathbf{D}'. \tag{39}$$

Post-multiplication of (39) by $\mathbf{X}'\mathbf{X}$ and application of Lemma 5 shows that $\mathbf{B}_{11}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Furthermore, using (37), (39), and Lemmas 5 and 6 gives

$$(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})[\mathbf{B}_{11} + \mathbf{D}(\mathbf{D}'\mathbf{H}'\mathbf{H}\mathbf{D})^{-1}\mathbf{D}'] = \mathbf{I}. \tag{40}$$

From (40),

$$(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1} = \mathbf{B}_{11} + \mathbf{D}(\mathbf{D}'\mathbf{H}'\mathbf{H}\mathbf{D})^{-1}\mathbf{D}'. \tag{41}$$

Since from Lemma 5, $\mathbf{D}$ is such that $\mathbf{X}\mathbf{D} = \mathbf{0}$ we have that

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}\mathbf{B}_{11}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$$

since $\mathbf{B}_{11}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. This completes the proof. ■

It can be shown that $\mathbf{B}_{11}$ satisfies the second of Penrose conditions and is thus a reflexive generalized inverse of $\mathbf{X}'\mathbf{X}$. However, $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}$ only satisfies the first Penrose condition. Neither generalized inverse satisfies conditions (iii) or (iv).

John (1964) refers to Graybill (1961, p. 292) and to Kempthorne (1952, p. 79) in discussing $\mathbf{B}_{11}$ and to Plackett (1960, p. 41) and Scheffe (1959, p. 19) in discussing $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}$, in terms of defining generalized inverses of $\mathbf{X}'\mathbf{X}$ as being matrices $\mathbf{G}$ for which $\mathbf{b} = \mathbf{G}\mathbf{X}'\mathbf{y}$ is a solution of $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$. By Theorem 1, they then satisfy condition (i), as has just been shown. Rayner and Pringle (1967) also discuss these results, indicating that $\mathbf{D}$ of the previous discussion may be taken as $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$. This, as Chipman (1964) shows, means that $\mathbf{H}\mathbf{D} = \mathbf{I}$ and so (39) becomes

$$\mathbf{X}'\mathbf{X}\mathbf{B}_{11} = \mathbf{I} - \mathbf{H}'\mathbf{H}(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}, \tag{42}$$

a simplified form of Rayner and Pringle's equation (7). The relationship between the two generalized inverses of $\mathbf{X}'\mathbf{X}$ shown in (35) is therefore that indicated in (42). Also note that Lemma 6 is equivalent to Theorem 3 of Scheffe (1959, p. 17).

## 6. ARBITRARINESS IN A GENERALIZED INVERSE

The existence of many generalized inverses $\mathbf{G}$ that satisfy $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$ has been emphasized. We examine here the nature of the arbitrariness of such generalized inverses as discussed in Urquhart (1969a). We need some results about the rank of the matrix. These are contained in Lemmas 7–9.

**Lemma 7** A matrix of full-row rank $r$ can be written as the product of matrices, one being of the form $\begin{bmatrix}\mathbf{I}_r & \mathbf{S}\end{bmatrix}$ for some matrix $\mathbf{S}$ of $r$ rows.

*Proof.* Suppose $\mathbf{B}_{r\times q}$ has full-row rank $r$ and contains an $r \times r$ non-singular minor, $\mathbf{M}$, say. Then, for some matrix $\mathbf{L}$ and some permutation matrix $\mathbf{Q}$ (see the paragraph just before (9)), we have $\mathbf{BQ} = \begin{bmatrix}\mathbf{M} & \mathbf{L}\end{bmatrix}$, so that

$$\mathbf{B} = \mathbf{M}\begin{bmatrix}\mathbf{I} & \mathbf{M}^{-1}\mathbf{L}\end{bmatrix}\mathbf{Q}^{-1} = \mathbf{M}\begin{bmatrix}\mathbf{I} & \mathbf{S}\end{bmatrix}\mathbf{Q}^{-1} \text{ for } \mathbf{S} = \mathbf{M}^{-1}\mathbf{L}.$$

■

**Lemma 8** $\mathbf{I} + \mathbf{KK}'$ has full rank for any non-null matrix $\mathbf{K}$.

*Proof.* Assume that $\mathbf{I} + \mathbf{KK}'$ does not have full rank. Then its columns are not LIN and there exists a non-null vector $\mathbf{u}$ such that

$$(\mathbf{I} + \mathbf{KK}')\mathbf{u} = \mathbf{0}, \text{ so that } \mathbf{u}'(\mathbf{I} + \mathbf{KK}')\mathbf{u} = \mathbf{u}'\mathbf{u} + \mathbf{u}'\mathbf{K}(\mathbf{u}'\mathbf{K})' = \mathbf{0}.$$

However, $\mathbf{u}'\mathbf{u}$ and $\mathbf{u}'\mathbf{K}(\mathbf{u}'\mathbf{K})'$ are both sums of squares of real numbers. Hence, their sum is zero only if their elements are zero, that is, only if $\mathbf{u} = \mathbf{0}$. This contradicts the assumption. Therefore, $\mathbf{I} + \mathbf{KK}'$ has full rank. ■

**Lemma 9** When $\mathbf{B}$ has full row rank, $\mathbf{BB}'$ is non-singular.

*Proof.* As in Lemma 7 write $\mathbf{B} = \mathbf{M}\begin{bmatrix}\mathbf{I} & \mathbf{S}\end{bmatrix}\mathbf{Q}^{-1}$ where $\mathbf{M}^{-1}$ exists. Then because $\mathbf{Q}$ is a permutation matrix and thus orthogonal $\mathbf{BB}' = \mathbf{M}(\mathbf{I} + \mathbf{SS}')\mathbf{M}'$. By virtue of Lemma 8 and the existence of $\mathbf{M}^{-1}$, $\mathbf{BB}'$ is non-singular. ■

**Corollary 9.1** When $\mathbf{B}$ has full-column rank, $\mathbf{BB}'$ is non-singular.

*Proof.* When $\mathbf{B}$ has full column rank $\mathbf{B}'$ has full-row rank. Now

$$\mathbf{BB}' = (\mathbf{B}'\mathbf{B})' = (\mathbf{B}'(\mathbf{B}')')'$$

From Lemma 9, $\mathbf{B}'(\mathbf{B}')'$ is non-singular and so is its transpose. ■

Consider now a matrix $\mathbf{A}_{p\times q}$ of rank $r$, less than both $p$ and $q$. The matrix $\mathbf{A}$ contains at least one non-singular minor of order $r$. We will assume that this is the leading minor. There is no loss of generality in this assumption because, if it is not true, the algorithm of Section 1b will always yield a generalized inverse of $\mathbf{A}$. This generalized inverse will come from a generalized inverse of $\mathbf{B} = \mathbf{RAS}$ where $\mathbf{R}$ and $\mathbf{S}$ are permutation matrices so that $\mathbf{B}$ has a non-singular $r \times r$ leading minor. We therefore confine the discussion of inverses of $\mathbf{A}$ to the case where its leading $r \times r$ minor is non-singular. Accordingly, $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \begin{bmatrix}(\mathbf{A}_{11})_{r\times r} & (\mathbf{A}_{12})_{r\times(q-r)} \\ (\mathbf{A}_{21})_{(p-r)\times r} & (\mathbf{A}_{22})_{(p-r)\times(q-r)}\end{bmatrix}. \tag{43}$$

Then since $\mathbf{A}_{11}^{-1}$ exists, $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} \end{bmatrix} \mathbf{A}_{11} \begin{bmatrix} \mathbf{I} & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix} = \mathbf{L}\mathbf{A}_{11}\mathbf{M} \tag{44}$$

with $\mathbf{L} = \begin{bmatrix} \mathbf{I} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} \end{bmatrix}$ and $\mathbf{M} = \begin{bmatrix} \mathbf{I} & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix}$. From Lemma 4, $\mathbf{L}$ has full-column rank and $\mathbf{M}$ has full-row rank. Lemma 9 shows that both $(\mathbf{L}'\mathbf{L})^{-1}$ and $(\mathbf{M}'\mathbf{M})^{-1}$ exist.

The arbitrariness in a generalized inverse of $\mathbf{A}$ is investigated by means of this partitioning. Thus, on substituting (44) into $\mathbf{AGA} = \mathbf{A}$, we get

$$\mathbf{L}\mathbf{A}_{11}\mathbf{M}\mathbf{G}\mathbf{L}\mathbf{A}_{11}\mathbf{M} = \mathbf{L}\mathbf{A}_{11}\mathbf{M}. \tag{45}$$

Pre-multiplication by $\mathbf{A}_{11}^{-1}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'$ and post-multiplication by $\mathbf{M}'(\mathbf{M}'\mathbf{M})^{-1}\mathbf{A}_{11}^{-1}$ then gives

$$\mathbf{M}\mathbf{G}\mathbf{L} = \mathbf{A}_{11}^{-1}. \tag{46}$$

Whatever the generalized inverse is, suppose it is partitioned as

$$\mathbf{G} = \begin{bmatrix} (\mathbf{G}_{11})_{r\times r} & (\mathbf{G}_{12})_{r\times(p-r)} \\ (\mathbf{G}_{21})_{(q-r)\times r} & (\mathbf{G}_{22})_{(q-r)\times(p-r)} \end{bmatrix} \tag{47}$$

of order $q \times p$, conformable for multiplication with $\mathbf{A}$. Then substituting (47) and (44) into (46) gives

$$\mathbf{G}_{11} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{G}_{21} + \mathbf{G}_{12}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{G}_{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} = \mathbf{A}_{11}^{-1}. \tag{48}$$

This is true whatever the generalized inverse may be. Therefore, on substituting from (48) for $\mathbf{G}_{11}$, we have

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{G}_{21} - \mathbf{G}_{12}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{G}_{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} \tag{49}$$

as a generalized inverse of $\mathbf{A}$ for any matrices $\mathbf{G}_{12}$, $\mathbf{G}_{21}$, and $\mathbf{G}_{22}$ of appropriate order. Thus, the arbitrariness of a generalized inverse is characterized.

**Example 13** Illustration of the Characterization in (49)

Let $\mathbf{A} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$ and $\mathbf{G} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}$. This generalized inverse only satisfies Penrose condition (i). Partition $\mathbf{A}$ so that $\mathbf{A}_{11} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$, $\mathbf{A}_{12} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$,

$\mathbf{A}_{21} = \begin{bmatrix} 2 & 0 \end{bmatrix}$, and $\mathbf{A}_{22} = [2]$. Also $\mathbf{G}_{11} = \begin{bmatrix} \frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}$, $\mathbf{G}_{12} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{G}_{21} = \begin{bmatrix} -\frac{1}{4} & 0 \end{bmatrix}$, and $\mathbf{G}_{22} = \begin{bmatrix} \frac{1}{2} \end{bmatrix}$. Now $\mathbf{A}_{11}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$. Using Formula 49, we can see that

$$\mathbf{G}_{11} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{4} & 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

□

Certain consequences of (49) can be noted.

1. By making $\mathbf{G}_{12}, \mathbf{G}_{21},$ and $\mathbf{G}_{22}$ null, $\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, a form discussed earlier.
2. When $\mathbf{A}$ is symmetric, $\mathbf{G}$ is not necessarily symmetric. Only when $\mathbf{G}_{12} = \mathbf{G}_{21}'$ and $\mathbf{G}_{22}$ is symmetric will $\mathbf{G}$ be symmetric.
3. When $p \geq q$, $\mathbf{G}$ can have full row rank $q$ even if $r < q$. For example, if $\mathbf{G}_{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{G}_{22}$, $\mathbf{G}_{21} = \mathbf{0}$ and $\mathbf{G}_{22}$ has full row rank the rank of $\mathbf{G}$ can exceed the rank of $\mathbf{A}$. In particular, this means that singular matrices can have non-singular generalized inverses.

The arbitrariness evident in (49) prompts investigating the relationship of one generalized inverse to another. It is simple. If $\mathbf{G}_1$ is a generalized inverse of **A,** then so is

$$\mathbf{G} = \mathbf{G}_1\mathbf{A}\mathbf{G}_1 + (\mathbf{I} - \mathbf{G}_1\mathbf{A})\mathbf{X} + \mathbf{Y}(\mathbf{I} - \mathbf{A}\mathbf{G}_1) \tag{50}$$

for any $\mathbf{X}$ and $\mathbf{Y}$. Pre- and post-multiplication of (50) by $\mathbf{A}$ shows that this is so.

The importance of (50) is that it provides a method of generating all generalized inverses of $\mathbf{A}$. They can all be put in the form of (50). To see this, we need only show that for some other generalized inverse $\mathbf{G}_2$ that is different from $\mathbf{G}_1$, there exist values of $\mathbf{X}$ and $\mathbf{Y}$ giving $\mathbf{G} = \mathbf{G}_2$. Putting $\mathbf{X} = \mathbf{G}_2$ and $\mathbf{Y} = \mathbf{G}_1\mathbf{A}\mathbf{G}_2$ achieves this.

The form of $\mathbf{G}$ in (50) is entirely compatible with the partitioned form given in (49). For if we take $\mathbf{G}_1 = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and partition $\mathbf{X}$ and $\mathbf{Y}$ in the same manner as $\mathbf{G}$, then (50) becomes

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{X}_{21} - \mathbf{Y}_{12}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{X}_{22} + \mathbf{Y}_{12} \\ \mathbf{X}_{21} - \mathbf{Y}_{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{X}_{22} + \mathbf{Y}_{22} \end{bmatrix}. \tag{51}$$

This characterizes the arbitrariness even more specifically than does (49). Thus, for the four sub-matrices of $\mathbf{G}$ shown in (47) we have

| Sub-matrix | Source of Arbitrariness |
|---|---|
| $\mathbf{G}_{11}$ | $\mathbf{X}_{21}$ and $\mathbf{Y}_{12}$ |
| $\mathbf{G}_{12}$ | $\mathrm{X}_{22}$ and $\mathrm{Y}_{12}$ |
| $\mathbf{G}_{21}$ | $\mathbf{X}_{21}$ and $\mathbf{Y}_{22}$ |
| $\mathbf{G}_{22}$ | $\mathbf{X}_{22}$ and $\mathbf{Y}_{22}$ |

This means that in the partitioning of

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix}$$

implicit in (50), the first set of rows in the partitioning of $\mathbf{X}$ does not enter into $\mathbf{G}$, and neither does the first set of columns of $\mathbf{Y}$.

It has been shown earlier (Theorem 3) that all solutions to $\mathbf{Ax} = \mathbf{y}$ can be generated from $\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{GA} - \mathbf{I})\mathbf{z}$, where $\mathbf{z}$ is the infinite set of arbitrary vectors of order $q$. We now show that all solutions can be generated from $\tilde{\mathbf{x}} = \mathbf{Gy}$ where $\mathbf{G}$ is the infinite set of generalized inverses indicated in (50). First, a Lemma is needed.

**Lemma 10** If $\mathbf{z}_{q\times 1}$ is arbitrary and $\mathbf{y}_{p\times 1}$ is known and non-null, there exists an arbitrary matrix $\mathbf{X}$ such that $\mathbf{z} = \mathbf{Xy}$.

*Proof.* Since $\mathbf{y} \neq \mathbf{0}$ at least one element $y_k$ say, will be non-zero. Writing $\mathbf{z} = \{z_j\}$ and $\mathbf{X} = \{x_{ij}\}$ for $i = 1,\ldots, q$ and $j = 1,\ldots, p$, let $x_{ij} = z_i/y_k$ for $j = k$ and $x_{ij} = 0$ otherwise. Then $\mathbf{Xy} = \mathbf{z}$ and $\mathbf{X}$ is arbitrary. ■

We use this lemma to prove the theorem on generating solutions.

**Theorem 11** For all possible generalized inverses $\mathbf{G}$ of $\mathbf{A}$, $\tilde{\mathbf{x}} = \mathbf{Gy}$ generates all solutions to the consistent equations $\mathbf{Ax} = \mathbf{y}$.

*Proof.* For the generalized inverse $\mathbf{G}_1$, solutions to $\mathbf{Ax} = \mathbf{y}$ are $\tilde{\mathbf{x}} = \mathbf{G}_1\mathbf{y} + (\mathbf{G}_1\mathbf{A} - \mathbf{I})\mathbf{z}$ where $\mathbf{z}$ is arbitrary. Let $\mathbf{z} = -\mathbf{Xy}$ for some arbitrary $\mathbf{X}$**.** Then

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{G}_1\mathbf{y} - (\mathbf{G}_1\mathbf{A} - \mathbf{I})\mathbf{Xy} \\ &= \mathbf{G}_1\mathbf{y} - \mathbf{G}_1\mathbf{A}\mathbf{G}_1\mathbf{y} + \mathbf{G}_1\mathbf{A}\mathbf{G}_1\mathbf{y} + (\mathbf{I} - \mathbf{G}_1\mathbf{A})\mathbf{Xy} \\ &= [\mathbf{G}_1\mathbf{A}\mathbf{G}_1 + (\mathbf{I} - \mathbf{G}_1\mathbf{A})\mathbf{X} + \mathbf{G}_1(\mathbf{I} - \mathbf{A}\mathbf{G}_1)\mathbf{y} \\ &= \mathbf{Gy}, \end{aligned}$$

where $\mathbf{G}$ is exactly the form given in (50) using $\mathbf{G}_1$ for $\mathbf{Y}$. ■

In Theorem 9 (iii), we showed how to represent any generalized inverse of $\mathbf{X}'\mathbf{X}$ in terms of the Moore–Penrose inverse. Theorem 12 shows how to do this for a generalized inverse of any matrix **A.**

**Theorem 12** Let **G** be any generalized inverse of **A.** Then

(i) $\mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{S}'\mathbf{S} = \mathbf{A}^+$.

(ii) $\mathbf{G} = \mathbf{X}^+ + \mathbf{U}\mathbf{C}_1\mathbf{T} + \mathbf{V}\mathbf{C}_2\mathbf{S} + \mathbf{V}\mathbf{C}_3\mathbf{T} = \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}^{-1/2} & \mathbf{C}_1 \\ \mathbf{C}_2 & \mathbf{C}_3\end{bmatrix}\begin{bmatrix}\mathbf{S} \\ \mathbf{T}\end{bmatrix}$,
where $\mathbf{C}_1 = \mathbf{V}'\mathbf{G}\mathbf{S}'$, $\mathbf{C}_2 = \mathbf{U}'\mathbf{G}\mathbf{T}'$ and $\mathbf{C}_3 = \mathbf{V}'\mathbf{G}\mathbf{T}'$.

*Proof.* (i) Since $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$, we have, using the singular value decomposition of **A**,

$$\mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}'\mathbf{G}\mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{S}'\mathbf{\Lambda}^{1/2}\mathbf{U}'. \tag{52}$$

Pre-multiply (52) by $\mathbf{\Lambda}^{-1/2}\mathbf{S}$ and post-multiply by $\mathbf{U}\mathbf{\Lambda}^{-1/2}$. Then we get

$$\mathbf{U}'\mathbf{G}\mathbf{S}' = \mathbf{\Lambda}^{-1/2}. \tag{53}$$

Pre-multiply (53) by **U** and post-multiply by **S** to obtain

$$\mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{S}'\mathbf{S} = \mathbf{A}^+.$$

(ii) Notice that

$$\begin{aligned}\mathbf{G} &= (\mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}')\mathbf{G}(\mathbf{S}'\mathbf{S} + \mathbf{T}'\mathbf{T}) \\ &= \mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{S}'\mathbf{S} + \mathbf{V}\mathbf{V}'\mathbf{G}\mathbf{S}'\mathbf{S} + \mathbf{U}\mathbf{U}'\mathbf{G}\mathbf{T}'\mathbf{T} + \mathbf{V}\mathbf{V}'\mathbf{G}\mathbf{T}'\mathbf{T} \\ &= \mathbf{A}^+ + \mathbf{V}\mathbf{C}_1\mathbf{S} + \mathbf{U}\mathbf{C}_2\mathbf{T} + \mathbf{V}\mathbf{C}_3\mathbf{T}.\end{aligned}$$

■

## 7. OTHER RESULTS

Procedures for inverting partitioned matrices are well-known (e.g., Section 8.7 of Searle (1966), Section 3 of Gruber (2014)). In particular, the inverse of the partitioned full-rank symmetric matrix

$$\mathbf{M} = \begin{bmatrix}\mathbf{X}' \\ \mathbf{Z}'\end{bmatrix}\begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D}\end{bmatrix}, \tag{54}$$

say, can for

$$\mathbf{W} = (\mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} = [\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}],$$

be written as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{W}\mathbf{B}'\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{W} \\ -\mathbf{W}\mathbf{B}'\mathbf{A}^{-1} & \mathbf{W} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{I} \end{bmatrix} \mathbf{W} \begin{bmatrix} -\mathbf{B}'\mathbf{A}^{-1} & \mathbf{I} \end{bmatrix}. \tag{55}$$

The analogy for (55) for generalized inverses, when $\mathbf{M}$ is symmetric but singular, has been derived by Rhode (1965). In defining $\mathbf{A}^-$ and $\mathbf{Q}^-$ as generalized inverses of $\mathbf{A}$ and $\mathbf{Q}$, respectively, where $\mathbf{Q} = \mathbf{D} - \mathbf{B}'\mathbf{A}^-\mathbf{B}$, then a generalized inverse of $\mathbf{M}$ is

$$\mathbf{M}^- = \begin{bmatrix} \mathbf{A}^- + \mathbf{A}^-\mathbf{B}\mathbf{Q}^-\mathbf{B}'\mathbf{A}^- & -\mathbf{A}^-\mathbf{B}\mathbf{Q}^- \\ -\mathbf{Q}^-\mathbf{B}'\mathbf{A}^- & \mathbf{Q}^- \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{A}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}^-\mathbf{B} \\ \mathbf{I} \end{bmatrix} \mathbf{Q}^- \begin{bmatrix} -\mathbf{B}'\mathbf{A}^- & \mathbf{I} \end{bmatrix}. \tag{56}$$

It is to be emphasized that the generalized inverses referred to here are just as have been defined throughout, namely satisfying only the first of Penrose's four conditions. (In showing that $\mathbf{M}\mathbf{M}^-\mathbf{M} = \mathbf{M}$, considerable use is made of Theorem 7.)

**Example 14** A Generalized Inverse of a Partitioned Matrix

Consider the matrix with the partitioning,

$$\mathbf{M} = \begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \mathbf{D}, \mathbf{B} = \mathbf{B}' = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

A generalized inverse of $\mathbf{A}$ is

$$\mathbf{A}^- = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix},$$
$$\mathbf{Q} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & \frac{3}{2} \\ \frac{3}{2} & \frac{3}{2} \end{bmatrix} \text{ and}$$
$$\mathbf{Q}^- = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then,

$$\mathbf{M}^{-} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2}{3} & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

We could have used different generalized inverses for **A** and **Q**. If we had done so, we would get a valid but different generalized inverse for **M**.

The regular inverse of the product **AB** is $\mathbf{B}^{-1}\mathbf{A}^{-1}$. However, there is no analogous result for generalized inverses. When one matrix is non-singular, **B**, say, Rohde (1964) indicates that $\mathbf{B}^{-1}\mathbf{A}^{-}$ is a generalized inverse of **AB**. Greville (1966) considers the situation for unique generalized inverses $\mathbf{A}^{(\mathrm{p})}$ and $\mathbf{B}^{(\mathrm{p})}$, and gives five separate conditions under which $(\mathbf{AB})^{(\mathrm{p})} = \mathbf{B}^{(\mathrm{p})}\mathbf{A}^{(\mathrm{p})}$. However, one would hope for conditions less complex that those of Greville for generalized inverses $\mathbf{A}^{-}$and $\mathbf{B}^{-}$ satisfying just the first of Penrose's conditions. What can be shown is that $\mathbf{B}^{-}\mathbf{A}^{-}$ is a generalized inverse of **AB** if and only if $\mathbf{A}^{-}\mathbf{ABB}^{-}$ is idempotent. Furthermore, when the product **AB** is itself idempotent, it has **AB**, $\mathbf{AA}^{-}$, $\mathbf{BB}^{-}$, and $\mathbf{B}^{-}\mathbf{BAA}^{-}$ as generalized inverses. Other problems of interest are the generalized inverse of $\mathbf{A}^{k}$ in terms of that of **A**, for integer $k$, and the generalized inverse of $\mathbf{XX}'$ in terms of that of $\mathbf{X}'\mathbf{X}$.

## 8. EXERCISES

**1** Reduce the matrices

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 1 & -1 \\ 5 & 8 & 0 & 1 \\ 1 & 2 & -2 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 4 & 5 & 6 & 2 \\ 7 & 8 & 10 & 7 \\ 2 & 1 & 1 & 6 \end{bmatrix}$$

to diagonal form and find a generalized inverse of each.

**2** Find generalized inverses of **A** and **B** in Exercise 1 by inverting non-singular minors.

**3** For **A** and **B** of Exercise 1, find general solutions to

**(a)** $\mathbf{AX} = \begin{bmatrix} -1 \\ -13 \\ 11 \end{bmatrix}$

**(b)** $\mathbf{Bx} = \begin{bmatrix} 14 \\ 23 \\ 32 \\ -5 \end{bmatrix}$

**4** Find the Penrose inverse of $\begin{bmatrix} 1 & 0 & 2 \\ 2 & -1 & 5 \\ 0 & 1 & -1 \\ 1 & 3 & -1 \end{bmatrix}$.

**5** Which of the remaining axioms for a Moore–Penrose inverse are satisfied by the generalized inverse in Example 2?

**6** **(a)** Using the Algorithm in Section 1b, find generalized inverses of

$$\mathbf{A}_1 = \begin{bmatrix} 4 & 1 & 2 & 0 \\ 1 & 1 & 5 & 15 \\ 3 & 1 & 3 & 5 \end{bmatrix}$$

derived from inverting the $2 \times 2$ minors

$$\mathbf{M}_1 = \begin{bmatrix} 1 & 5 \\ 1 & 3 \end{bmatrix}, \mathbf{M}_2 = \begin{bmatrix} 1 & 15 \\ 1 & 5 \end{bmatrix}, \text{ and } \mathbf{M}_3 = \begin{bmatrix} 4 & 0 \\ 3 & 5 \end{bmatrix}.$$

**(b)** Using the Algorithm in Section 1b find a generalized inverse of

$$\mathbf{A}_2 = \begin{bmatrix} 2 & 2 & 6 \\ 2 & 3 & 8 \\ 6 & 8 & 22 \end{bmatrix}$$

derived from inverting the minor

$$\mathbf{M} = \begin{bmatrix} 3 & 8 \\ 8 & 22 \end{bmatrix}.$$

**7** Let

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

**(a)** Find the Moore–Penrose inverse of **A.**

**(b)** Classify the following generalized inverses of **A** as named in Table 1.1 by determining which of the Penrose conditions are satisfied.

**(i)** $\begin{bmatrix} \frac{3}{4} & \frac{9}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$

**(ii)** $\begin{bmatrix} \frac{5}{4} & \frac{3}{4} \\ \frac{7}{4} & \frac{1}{4} \end{bmatrix}$

**(iii)** $\begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ 1 & \frac{1}{2} \end{bmatrix}$

**(iv)** $\begin{bmatrix} \frac{3}{4} & \frac{5}{4} \\ \frac{5}{4} & \frac{3}{4} \end{bmatrix}$

**8** Given $\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$

Find

**(a)** A minimum norm generalized inverse of **X**

**(b)** A least-square generalized inverse of **X**

**(c)** The Moore–Penrose inverse of **X**

**9** Find a generalized inverse of each of the following matrices.

**(a)** **PAQ**, when **P** and **Q** are non-singular

**(b)** **GA**, when **G** is a generalized inverse of **A**

**(c)** $k\mathbf{A}$, where $k$ is a scalar

**(d)** **ABA,** when **ABA** is idempotent

**(e)** **J**, when **J** is square with every element unity

**10** What kinds of matrices

**(a)** are their own generalized inverses?

**(b)** have transposes as a generalized inverse?

**(c)** have an identity matrix as a generalized inverse?

**(d)** have every matrix of order $p \times q$ for a generalized inverse?

**(e)** have only non-singular generalized inverses?

**11** Explain why the equations (a) $\mathbf{Ax} = \mathbf{0}$ and (b) $\mathbf{X'Xb} = \mathbf{X'y}$ are always consistent.

**12** If $\mathbf{z} = (\mathbf{G} - \mathbf{F})\mathbf{y} + (\mathbf{I} - \mathbf{FA})\mathbf{w}$, where **G** and **F** are generalized inverses of **A**, show that the solution $\tilde{\mathbf{x}} = \mathbf{Gy} + (\mathbf{GA} - \mathbf{I})\mathbf{z}$ to $\mathbf{Ax} = \mathbf{y}$ reduces to $\tilde{\mathbf{x}} = \mathbf{Fy} + (\mathbf{FA} - \mathbf{I})\mathbf{w}$.

**13** If $\mathbf{Ax} = \mathbf{y}$ are consistent equations, and $\mathbf{F}$ and $\mathbf{G}$ are generalized inverses of $\mathbf{A}$, find in simplest form, a solution for $\mathbf{w}$ to the equations

$$(\mathbf{I} - \mathbf{GA})\mathbf{w} = (\mathbf{F} - \mathbf{G})\mathbf{y} + (\mathbf{FA} - \mathbf{I})\mathbf{z}.$$

**14** **(a)** If $\mathbf{A}$ has full-column rank**,** show that its generalized inverses are also left inverses satisfying the first three Penrose conditions.

**(b)** If $\mathbf{A}$ has full-row rank**,** show that its generalized inverses are also right inverses satisfying the first, second, and fourth Penrose conditions.

**15** Show that (29) reduces to (27).

**16** Give an example of a singular matrix that has a non-singular generalized inverse.

**17** Prove that $\mathbf{B}^{-}\mathbf{A}^{-}$ is a generalized inverse of $\mathbf{AB}$ if and only if $\mathbf{A}^{-}\mathbf{ABB}^{-}$ is idempotent.

**18** Show that the rank of a generalized inverse of $\mathbf{A}$ does not necessarily have the same rank as $\mathbf{A}$ and that it is the same if and only if it has a reflexive generalized inverse. See Rhode (1966), also see Ben-Israel and Greville (2003), and Harville (2008).

**19** When $\mathbf{PAQ} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ with $\mathbf{P}$ and $\mathbf{Q}$ non-singular show that $\mathbf{G} = \mathbf{Q}\begin{bmatrix} \mathbf{D}^{-1} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix}\mathbf{P}$ is a generalized inverse of $\mathbf{A}$. Under what conditions does $\mathbf{GAG} = \mathbf{G}$? Use $\mathbf{G}$ to answer Exercise 15.

**20** Using $\mathbf{AGA} = \mathbf{A}$

**(a)** Find a generalized inverse of $\mathbf{AB}$ where $\mathbf{B}$ is orthogonal.

**(b)** Find a generalized inverse of $\mathbf{LA}$ where $\mathbf{A}$ is non-singular.

**21** What is the Penrose inverse of a symmetric idempotent matrix?

**22** If $\mathbf{G}$ is a generalized inverse of $\mathbf{A}_{p\times q}$, show that $\mathbf{G} + \mathbf{Z} - \mathbf{GAZAG}$ generates

**(a)** all generalized inverses of $\mathbf{A}$, and

**(b)** all solutions to consistent equations $\mathbf{Ax} = \mathbf{y}$ as $\mathbf{Z}$ ranges over all matrices of order $q \times p$.

**23** Show that the generalized inverse of $\mathbf{X}$ that was derived in Theorem 12

$$\mathbf{G} = \mathbf{X}^{+} + \mathbf{UC}_1\mathbf{T} + \mathbf{VC}_2\mathbf{S} + \mathbf{VC}_3\mathbf{T} = \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix}\begin{bmatrix} \Lambda^{-1/2} & \mathbf{C}_1 \\ \mathbf{C}_2 & \mathbf{C}_3 \end{bmatrix}\begin{bmatrix} \mathbf{S} \\ \mathbf{T} \end{bmatrix}$$

**(a)** Satisfies Penrose condition (ii)(is reflexive) when $\mathbf{C}_3 = \mathbf{C}_2\Lambda^{1/2}\mathbf{C}_1$;

**(b)** Satisfies Penrose condition (iii)(is minimum norm) when $\mathbf{C}_2 = 0$;

**(c)** Satisfies Penrose Condition (iv)(is a least-square generalized inverse when $\mathbf{C}_1 = 0$.)

**24** Show that $\mathbf{M} = (\mathbf{X}'\mathbf{X})^+\mathbf{X}'$ and $\mathbf{W} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^+$ are expressions for the Moore–Penrose inverse of $\mathbf{X}$

**(i)** by direct verification of the four Penrose conditions.

**(ii)** using the singular value decomposition.

**25** Show that if $\mathbf{N}$ is a non-singular matrix, then $(\mathbf{UNU}')^+ = \mathbf{UN}^{-1}\mathbf{U}'$.

**26** Show that if $\mathbf{P}$ is an orthogonal matrix, $(\mathbf{PAP}')^+ = \mathbf{PA}^+\mathbf{P}'$.

**27** Show that

**(a)** $\mathbf{X}^+(\mathbf{X}')^+ = (\mathbf{X}'\mathbf{X})^+$;

**(b)** $(\mathbf{X}')^+\mathbf{X}^+ = (\mathbf{XX}')^+$.

**28** Show that the generalized inverses that would be produced by the algorithms in Sections 1a and 1b are reflexive.

**29** Show that $\mathbf{K}$ as defined in equation (30) satisfies the four Penrose axioms.

**30** Show that if $\mathbf{X}^-$ satisfies Penrose's condition (iv) then $\mathbf{b} = \mathbf{X}^-\mathbf{y}$ is a solution to $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$. [*Hint:* use Exercise 22 or Theorem 12.]

**31** Show that $\mathbf{M}^-$ of (56) is a generalized inverse of $\mathbf{M}$ in (54).

**32** If $\mathbf{P}_{m\times q}$ and $\mathbf{D}_{m\times m}$ have rank $m$ show that $\mathbf{D}^{-1} = \mathbf{P}(\mathbf{P}'\mathbf{DP})^-\mathbf{P}'$.

**33** Show by direct multiplication that

$$\mathbf{M}^- = \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{Z}'\mathbf{Z})^- \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ -(\mathbf{Z}'\mathbf{Z})^-\mathbf{Z}'\mathbf{X} \end{bmatrix} \mathbf{Q}^- \begin{bmatrix} \mathbf{I} & -\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^- \end{bmatrix},$$

where $\mathbf{Q} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^-\mathbf{Z}'\mathbf{X}$ is a generalized inverse of

$$\mathbf{M} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}.$$

# 2

# DISTRIBUTIONS AND QUADRATIC FORMS

## 1. INTRODUCTION

Analysis of variance techniques involve partitioning a total sum of squares into component sums of squares whose ratios (under appropriate distributional conditions) lead to $F$-statistics suitable for testing certain hypothesis. When discussing linear models generally, especially where unbalanced data (data having unequal subclass numbers) are concerned, it is convenient to think of sums of squares involved in this process as quadratic forms in the observations. In this context, we can establish very general theorems, for which familiar analysis of variance and the associated $F$-tests are then just special cases. An introductory outline[1] of the general procedure is easy to describe.

Suppose $\mathbf{y}_{n\times 1}$ is a vector of $n$ observations. The total sum of squares is $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2$. In an analysis of variance, the total sum of squares is partitioned into component sums of squares. Let $\mathbf{P}$ be an orthogonal matrix. Recall that an orthogonal matrix is one where

$$\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}. \tag{1}$$

[1] Kindly brought to the notice of S. R. Searle by D. L. Weeks.

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

Partition $\mathbf{P}$ row-wise into $k$ sub-matrices $\mathbf{P}_i$ of order $n_i \times n$ for $i = 1, 2, \ldots, k$ with $\sum_{i=1}^{k} n_i = n$; that is,

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \vdots \\ \mathbf{P}_k \end{bmatrix} \quad \text{and} \quad \mathbf{P}' = \begin{bmatrix} \mathbf{P}_1' & \mathbf{P}_2' & \cdots & \mathbf{P}_k' \end{bmatrix} \tag{2}$$

Then

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{P}'\mathbf{P}\mathbf{y} = \sum_{i=1}^{k} \mathbf{y}'\mathbf{P}_i'\mathbf{P}_i\mathbf{y}. \tag{3}$$

In this way, $\mathbf{y}'\mathbf{y}$ is partitioned into $k$ sums of squares

$$\mathbf{y}'\mathbf{P}_i'\mathbf{P}_i\mathbf{y} = \mathbf{z}_i'\mathbf{z}_i = \sum_{j=1}^{n_i} \mathbf{z}_{ij}^2 \quad \text{for} \quad i = 1, \ldots, n_i,$$

where

$$\mathbf{z}_i = \mathbf{P}_i\mathbf{y} = \{\mathbf{z}_{ij}\} \quad \text{for} \quad j = 1, 2, \ldots, n_i.$$

Each of these sums of squares corresponds to the lines in an analysis of variance table (with, as we shall see, degrees of freedom equal to the rank of $\mathbf{P}_i$), having $\mathbf{y}'\mathbf{y}$ as the total sum of squares. We can demonstrate the general nature of the results we shall develop in this chapter for the $k$ terms of $\mathbf{y}'\mathbf{P}_i'\mathbf{P}_i\mathbf{y}$ of equation (3). First, for example, in Corollary 2.1 of Theorem 2, we show that if elements of the $\mathbf{y}$ vector are normally and independently distributed with zero mean and variance $\sigma^2$, then $\mathbf{y}'\mathbf{A}\mathbf{y}/\sigma^2$, where $\mathbf{A}$ has rank $r$, has a $\chi^2$-distribution with $r$ degrees of freedom if and only if $\mathbf{A}$ is idempotent. This is just the property that the matrix $\mathbf{P}_i'\mathbf{P}_i$ has in equation (3). Observe that $\mathbf{P}_i'\mathbf{P}_i\mathbf{P}_i'\mathbf{P}_i = \mathbf{P}_i'(\mathbf{P}_i\mathbf{P}_i')\mathbf{P}_i = \mathbf{P}_i'\mathbf{I}\mathbf{P}_i = \mathbf{P}_i'\mathbf{P}_i$ because $\mathbf{P}'\mathbf{P} = \mathbf{I}$ in equation (1). Since each $\mathbf{P}_i'\mathbf{P}_i$ in equation (3) is idempotent, each term $\mathbf{y}'\mathbf{P}_i'\mathbf{P}_i\mathbf{y}/\sigma^2$ has a $\chi^2$-distribution. Second, in Theorem 6, we prove that when the elements of $\mathbf{y}$ are normally distributed as just described, $\mathbf{y}'\mathbf{A}\mathbf{y}$ and $\mathbf{y}'\mathbf{B}\mathbf{y}$ are independent if and only if $\mathbf{A}\mathbf{B} = \mathbf{0}$. This is also true for the terms in equation 3. If $i \neq j$, $\mathbf{P}_i\mathbf{P}_j' = 0$ from equations (1) and (2). Consequently,

$$\mathbf{P}_i'\mathbf{P}_i\mathbf{P}_j'\mathbf{P}_j = 0.$$

As a result, the terms in equation (3) are independent. Moreover, since they all have $\chi^2$-distributions, their ratios, suitably modified by degrees of freedom, can be $F$-distributions. In this way, tests of hypothesis may be established. We now give an illustrative example.

**Example 1 Development of an *F*-test** Corresponding to a vector of four observations consider the orthogonal matrix

$$\left[\begin{array}{cccc} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \\ \hline \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{3}{\sqrt{12}} \end{array}\right] = \left[\begin{array}{c} \mathbf{P}_1 \\ \hline \mathbf{P}_2 \end{array}\right] \tag{4}$$

partitioned as shown. The reader may verify, using equation (1) that $\mathbf{P}$ is orthogonal and that

$$\mathbf{P}_1\mathbf{y} = \frac{1}{\sqrt{4}}\sum_{i=1}^{4}\mathbf{y}_i = \frac{\sqrt{4}}{4}\sum_{i=1}^{4}\mathbf{y}_i = \sqrt{4}\bar{\mathbf{y}}.$$

From equation (3)

$$\mathbf{q}_1 = \mathbf{y}'\mathbf{P}_1'\mathbf{P}_1\mathbf{y} = 4\bar{\mathbf{y}}^2.$$

We also have that

$$\mathbf{q}_2 = \mathbf{y}'\mathbf{P}_2'\mathbf{P}_2\mathbf{y} = \sum_{i=1}^{4}\mathbf{y}_i^2 - 4\bar{\mathbf{y}}^2 = \sum_{i=1}^{4}(\mathbf{y}_i - \bar{\mathbf{y}})^2.$$

Therefore, when the elements of $\mathbf{y}$ are normally and independently distributed with mean zero and unit variance $\mathbf{q}_1$ and $\mathbf{q}_2$, each has $\chi^2$-distributions. From the orthogonality of $\mathbf{P}$, it follows that $\mathbf{P}_1'\mathbf{P}_2 = 0$. Thus, $\mathbf{q}_1$ and $\mathbf{q}_2$ are also distributed independently. As a result, the statistic

$$F = \frac{4\bar{\mathbf{y}}^2/1}{\left(\sum_{i=1}^{4}\mathbf{y}_i^2 - 4\bar{\mathbf{y}}^2\right)/3}$$

provides an *F*-test for the hypothesis that the mean of the *y*-variable is zero. □

The matrix $\mathbf{P}$ in (4) is a fourth order Helmert matrix. We now give the general characteristics of an *n*th order Helmert matrix. We may write an *n*th order Helmert matrix as

$$\mathbf{H}_{n\times n} = \begin{bmatrix} \mathbf{h}' \\ \mathbf{H}_0 \end{bmatrix} \quad \begin{matrix} 1 \times n \\ (n-1)\times n. \end{matrix}$$

For the first row, we have

$$\mathbf{h}' = \frac{1}{\sqrt{n}} 1_n'$$

where

$$1_n' = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix},$$

a vector on $n$ ones. Now, $\mathbf{H}_0$ consists of the last $n-1$ rows with the $r$th row

$$\begin{bmatrix} \frac{1}{\sqrt{r(r+1)}} 1_r' & \frac{-r}{\sqrt{r(r+1)}} & 0_{(n-r-1)\times 1} \end{bmatrix} \quad \text{for} \quad r = 1, 2, \ldots, n-1.$$

We have that $\mathbf{H}_{n\times n}$ is an orthogonal matrix. Furthermore, $\mathbf{y}'\mathbf{h}\mathbf{h}'\mathbf{y} = n\bar{\mathbf{y}}^2$. Using mathematical induction, it is readily shown that $\mathbf{y}'\mathbf{H}_0'\mathbf{H}_0\mathbf{y} = \sum_{i=1}^n \mathbf{y}_i^2 - n\bar{\mathbf{y}}^2$. Further properties of Helmert matrices are available in Lancaster (1965), for example.

## 2. SYMMETRIC MATRICES

An expression of the form $\mathbf{x}'\mathbf{A}\mathbf{y}$ is called a bilinear form. It is a homogeneous second-degree function of the first degree in each of the $x$'s and $y$'s. For example,

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{y} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ -2 & 7 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\ &= 4x_1y_1 + 8x_1y_2 - 2x_2y_1 + 7x_2y_2. \end{aligned}$$

When $\mathbf{x}$ is used in place of $\mathbf{y}$, the expression becomes $\mathbf{x}'\mathbf{A}\mathbf{x}$. It is then called a quadratic form and is a quadratic function of the $x$'s. Then we have

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ -2 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 4x_1^2 + (8-2)x_1x_2 + 7x_2^2 \\ &= 4x_1^2 + (3+3)x_1x_2 + 7x_2^2 \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

In this way, we can write any quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ as $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{B}\mathbf{x}$ where $\mathbf{B} = \frac{1}{2}(\mathbf{A} + \mathbf{A}')$ is symmetric. While we can write every quadratic form as $\mathbf{x}'\mathbf{A}\mathbf{x}$ for an infinite number of matrices, we can only write $\mathbf{x}'\mathbf{B}\mathbf{x}$ one way for $\mathbf{B}$ symmetric.

For example

$$4x_1^2 + 6x_1x_2 + 7x_2^2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 3+a \\ 3-a & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

for any value of $a$. However, the matrix is symmetric only when $a = 0$. This means that for any particular quadratic form, there is only one unique matrix such that the quadratic form can be written as $\mathbf{x}'\mathbf{A}\mathbf{x}$ with $\mathbf{A}$ symmetric. Because of the uniqueness of this symmetric matrix, all further discussion of quadratic forms $\mathbf{x}'\mathbf{A}\mathbf{x}$ is confined to the case of $\mathbf{A}$ being symmetric.

## 3. POSITIVE DEFINITENESS

A property of some quadratic forms used repeatedly in what follows is that of positive definiteness. A quadratic form is said to be positive definite if it is positive for all values of $\mathbf{x}$ except $\mathbf{x} = \mathbf{0}$; that is, if

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ for all } \mathbf{x}, \text{ except } \mathbf{x} = 0,$$

then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite. And the corresponding (symmetric) matrix is also described as positive definite.

**Example 2 A Positive Definite Quadratic Form** Consider

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 13 & 1 & 4 \\ 1 & 13 & 4 \\ 4 & 4 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= 13x_1^2 + 13x_2^2 + 10x_3^2 + 2x_1x_2 + 8x_1x_3 + 8x_2x_3. \end{aligned}$$

Using the singular value decomposition of $\mathbf{A}$, we have

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 18 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \sqrt{6} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= 18y_1^2 + 12y_2^2 + 6y_3^2, \end{aligned}$$

where

$$y_1 = \frac{1}{\sqrt{3}}(x_1 + x_2 + x_3), \quad y_2 = \frac{1}{\sqrt{2}}(-x_1 + x_2), \quad y_3 = \frac{1}{\sqrt{6}}(-x_1 - x_2 + 2x_3). \quad \square$$

We have that $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ if $\mathbf{x} \neq \mathbf{0}$ because we have a sum of squares with positive coefficients and $y_1 = y_2 = y_3 = 0$ if and only if $x_1 = x_2 = x_3 = 0$. Thus $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite (p.d.).

A slight relaxation of the above definition concerns $\mathbf{x}'\mathbf{A}\mathbf{x}$ when its value is either positive or zero for all $\mathbf{x} \neq \mathbf{0}$. We define an $\mathbf{x}'\mathbf{A}\mathbf{x}$ of this nature as being *positive semi-definite* (p.s.d.) when

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ for all } \mathbf{x} \neq 0, \text{ with } \mathbf{x}'\mathbf{A}\mathbf{x} = 0 \text{ for at least one } \mathbf{x} \neq 0.$$

Under these conditions, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is a p.s.d. quadratic form and the corresponding symmetric matrix $\mathbf{A}$ is a p.s.d. matrix. This definition is widely accepted (see, for example, Graybill (1976) and Rao (1973)). For example, Scheffe (1959, p. 398) defines a p.s.d. matrix as one where $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$ without demanding that $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{0}$ for at least one non-null $\mathbf{x}$. This definition includes p.d. and p.s.d matrices. We will call such matrices non-negative definite (n.n.d.) matrices in keeping, for example, with Rao (1973, p. 35). Thus, n.n.d. matrices are either p.d. or p.s.d.

**Example 3 A Positive Semi-definite Quadratic Form** Consider

$$\begin{aligned}
\mathbf{x}'\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 12 & 0 & 6 \\ 0 & 12 & 6 \\ 6 & 6 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 18 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \sqrt{6} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
&= 18y_1^2 + 12y_2^2,
\end{aligned}$$

where $y_1$ and $y_2$ are as defined in Example 2. Clearly $\mathbf{x}'\mathbf{A}\mathbf{x}$ is non-negative definite. However $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ for $x_1 = 1$, $x_2 = 1$, $x_3 = -2$. Thus $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{A}$ are positive semi-definite but not positive definite. □

As another example, observe that

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2$$

is positive definite because it is zero only when $\mathbf{y} = \mathbf{0}$. However, notice that

$$\mathbf{y}'\mathbf{y} - n\bar{\mathbf{y}}^2 = \mathbf{y}'(\mathbf{I} - n^{-1}\mathbf{J}_n)\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2 - n\bar{\mathbf{y}}^2$$

is a positive semi-definite matrix that is not positive definite because it is zero when every element of $\mathbf{y}$ is the same.

We will now establish some results about non-negative definite matrices that we will use in the sequel. The first one about the determinants of the principal minors (the minors that contain part of the main diagonal) will be stated without proof.

**Lemma 1** The symmetric matrix $\mathbf{A}$ is positive definite if and only if its principal leading minors have positive determinants. (See Seelye (1958) for a proof.)

Positive definite matrices are non-singular. However, not all non-singular matrices are positive definite. For example, the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

is symmetric and non-singular but not positive definite or positive semi-definite.

The next Lemma is often useful.

**Lemma 2** For $\mathbf{P}$ non-singular, $\mathbf{P'AP}$ is or is not positive (semi-) definite according as $\mathbf{A}$ is or is not p.(s.)d.

*Proof.* Let $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$. Consider $\mathbf{x'Ax} = \mathbf{y'P'APy}$. When $\mathbf{x} = \mathbf{0}$, $\mathbf{y} = \mathbf{0}$ and $\mathbf{x'Ax} = \mathbf{y'P'APy} = 0$. For $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$, $\mathbf{y'P'APy} \geq 0$ according as $\mathbf{x'Ax} \geq 0$.

Hence $\mathbf{P'AP}$ is p.(s.)d. according as $\mathbf{A}$ is p.(s.)d. ■

Notice that in Examples 2 and 3 the eigenvalues of the matrices were non-negative.

This fact is generally true for positive semi-definite matrices as shown by the next lemma.

**Lemma 3** The eigenvalues of a positive (semi-) definite matrix are all positive (non-negative).

*Proof.* Suppose that $\lambda$ and $\mathbf{u} \neq \mathbf{0}$ are an eigenvalue and eigenvector of $\mathbf{A}$, respectively with $\mathbf{Au} = \lambda\mathbf{u}$. Then consider $\mathbf{u'Au} = \mathbf{u'}\lambda\mathbf{u} = \lambda\mathbf{u'u}$ for $\mathbf{u} \neq \mathbf{0}$. When $\mathbf{A}$ is p.d. $\mathbf{u'Au} > 0$. Thus, $\lambda\mathbf{u'u} > 0$ so $\lambda > 0$. Thus, we have that all eigenvalues of a p.d. matrix are positive. When A is p.s.d., $\mathbf{u'Au} \geq 0$ with $\mathbf{u'Au} = 0$ for at least one $\mathbf{u} \neq \mathbf{0}$. That means that $\lambda = 0$ for at least one $\mathbf{u} \neq \mathbf{0}$. As a result, all eigenvalues of a p.s.d. matrix are zero or positive. ■

Positive semi-definite matrices are singular. At least one of the eigenvalues is zero resulting in a zero determinant. However, not all singular matrices are positive semi-definite. All positive definite matrices are non-singular.

The next lemma gives a representation of a positive definite matrix in terms of a non-singular matrix.

**Lemma 4** A symmetric matrix is positive definite if and only if it can be written as $\mathbf{P}'\mathbf{P}$ for a non-singular $\mathbf{P}$.

*Proof.* If $\mathbf{A} = \mathbf{P}'\mathbf{P}$ for $\mathbf{P}$ non-singular, then $\mathbf{A}$ is symmetric and $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{P}'\mathbf{P}\mathbf{x}$, the sum of the squares of the elements of $\mathbf{P}\mathbf{x}$. Thus, $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{P}\mathbf{x} \neq 0$ and $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ for all $\mathbf{P}\mathbf{x} = \mathbf{0}$. However, since $\mathbf{P}$ is non-singular, $\mathbf{P}^{-1}$ exists so $\mathbf{x} = \mathbf{P}^{-1}\mathbf{P}\mathbf{x} = \mathbf{0}$. Thus, $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ only when $\mathbf{x} = \mathbf{0}$. Therefore $\mathbf{A}$ is p.d.

On the other hand, suppose that $\mathbf{A}$ is p.d. Since $\mathbf{A}$ is symmetric, there exists a matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{A}\mathbf{Q}'$ is a diagonal matrix with 0's and 1's in its diagonal. Since $\mathbf{A}$ is p.d., it has full rank. Then $\mathbf{Q}\mathbf{A}\mathbf{Q}' = \mathbf{I}$ and because $\mathbf{Q}$ is non-singular $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{Q}'^{-1}$ which is of the form $\mathbf{P}'\mathbf{P}$. ■

The matrix $\mathbf{A}'\mathbf{A}$ is always non-negative definite as will be shown in Lemma 5.

**Lemma 5** The matrix $\mathbf{A}'\mathbf{A}$ is positive definite when $\mathbf{A}$ has full-column rank and is positive semi-definite otherwise.

*Proof.* The quadratic form $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x}$ is equal to the sum of squares of the elements of $\mathbf{A}\mathbf{x}$. When $\mathbf{A}$ has full-column rank, $\mathbf{A}\mathbf{x} = \mathbf{0}$ only when $\mathbf{x} = \mathbf{0}$. Thus, $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{A}'\mathbf{A}$ is p.d. If $\mathbf{A}$ has less than full-column rank, $\mathbf{A}\mathbf{x} = \mathbf{0}$ for some $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{A}'\mathbf{A}$ is p.s.d. ■

**Corollary 1** The matrix $\mathbf{A}\mathbf{A}'$ is positive definite when $\mathbf{A}$ has full-row rank and is positive-semi-definite otherwise.

*Proof.* Let $\mathbf{B} = \mathbf{A}'$. The row rank of $\mathbf{A}$ is the column rank of $\mathbf{B}$. Then $\mathbf{A}\mathbf{A}' = \mathbf{B}'\mathbf{B}$. The result now follows from Lemma 5. ■

The next result concerns the sum of positive (semi-) definite matrices.

**Lemma 6** The sum of positive (semi) definite matrices is positive (semi-) definite.

*Proof.* Consider $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\left(\sum_i \mathbf{A}_i\right)\mathbf{x}$. ■

We now obtain a representation of a non-full rank symmetric matrix.

**Lemma 7** A symmetric matrix $\mathbf{A}$ of order $n$ and rank $r$ can be written as $\mathbf{L}\mathbf{L}'$
Where $\mathbf{L}$ is $n \times r$ of rank $r$, that is, $\mathbf{L}$ has full-column rank.

*Proof.* For some orthogonal $\mathbf{P}$

$$\mathbf{P}\mathbf{A}\mathbf{P}' = \begin{bmatrix} \mathbf{D}_r^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_r \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \end{bmatrix}$$

where $\mathbf{D}_r^2$ is diagonal of order r. Hence

$$\mathbf{A} = \mathbf{P}' \begin{bmatrix} \mathbf{D}_r \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \end{bmatrix} \mathbf{P} = \mathbf{L}\mathbf{L}',$$

where $\mathbf{L}' = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \end{bmatrix} \mathbf{P}$ of order $r \times n$ and full-row rank so that $\mathbf{L}$ is $n \times r$ of full-column rank. Notice that although $\mathbf{L}\mathbf{L}' = \mathbf{A}$, $\mathbf{L}'\mathbf{L} = \mathbf{D}_r^2$. In addition $\mathbf{L}'$ is real only when $\mathbf{A}$ is n.n.d., for only then are the non-zero elements guaranteed to be positive. ■

We now show that a matrix all of whose eigenvalues are zero or one is idempotent.

**Lemma 8** A symmetric matrix having eigenvalues zero and one is idempotent.

*Proof.* The singular value decomposition of such a matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}' \\ \mathbf{V}' \end{bmatrix} = \mathbf{U}\mathbf{U}'.$$

However, $\mathbf{A}^2 = \mathbf{U}\mathbf{U}'\mathbf{U}\mathbf{U}' = \mathbf{U}\mathbf{I}_r\mathbf{U}' = \mathbf{U}\mathbf{U}' = \mathbf{A}$ so $\mathbf{A}$ is idempotent. ■

Another result about when matrices are idempotent is the following.

**Lemma 9** If $\mathbf{A}$ and $\mathbf{V}$ are symmetric and $\mathbf{V}$ is positive definite, then if $\mathbf{A}\mathbf{V}$ has eigenvalues zero and one, it is idempotent.

*Proof.* The characteristic equation $\det(\mathbf{A}\mathbf{V} - \lambda\mathbf{V}) = 0$ has roots zero and one. Then by Lemma 4, $\mathbf{V} = \mathbf{P}'\mathbf{P}$ for some non-singular matrix $\mathbf{P}$. It follows that the equation $\det(\mathbf{P})\ \det(\mathbf{A}\mathbf{V} - \lambda\mathbf{I})\ \det(\mathbf{P}^{-1}) = 0$ has roots zero and one and, as a result, $\det(\mathbf{P}\mathbf{A}\mathbf{P}' - \lambda\mathbf{I}) = 0$ also has roots zero and one. By Lemma 8, $\mathbf{P}\mathbf{A}\mathbf{P}'$ is idempotent.

But $\mathbf{A}\mathbf{V} = \mathbf{A}\mathbf{P}'\mathbf{P}$. Then

$$(\mathbf{A}\mathbf{V})^2 = \mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V} = \mathbf{A}\mathbf{P}'\mathbf{P}\mathbf{A}\mathbf{P}'\mathbf{P} = \mathbf{P}^{-1}\mathbf{P}\mathbf{A}\mathbf{P}'\mathbf{P}\mathbf{A}\mathbf{P}'\mathbf{P} = \mathbf{P}^{-1}\mathbf{P}\mathbf{A}\mathbf{P}'\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{P}' = \mathbf{A}\mathbf{V}$$

and $\mathbf{A}\mathbf{V}$ is idempotent. ■

We are going to develop a criterion for comparing the "size" of two non-negative definite matrices. This will be called the Loewner ordering. It will be useful to us later for comparing the efficiency of estimators.

**Definition 1** Matrix $\mathbf{A}$ will be said to be greater than or equal to matrix $\mathbf{B}$ in the sense of the Loewner ordering if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Matrix $\mathbf{A}$ will be said to be greater than matrix $\mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive definite.

Two non-negative definite matrices may or may not be comparable under the Loewner ordering. The following Theorem will be useful for the comparison of estimators later on.

**Theorem 1** If $\mathbf{A}$ and $\mathbf{B}$ are positive definite matrices and if $\mathbf{B} - \mathbf{A}$ is a positive (semi-) definite matrix, then $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ is a positive (semi)-definite matrix.

*Proof.* The proof is based on that in Gruber (2014). It should be clear that $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ is symmetric. We first establish that if $\mathbf{I} - \mathbf{A}$ is a positive (semi)-definite, then $\mathbf{A}^{-1} - \mathbf{I}$ is positive (semi-) definite. The matrix $\mathbf{A}$ may be written as $\mathbf{A} = \mathbf{P}\Delta\mathbf{P}'$ where $\Delta$ is the diagonal matrix of eigenvalues ordered from highest to lowest and $\mathbf{P}$ is an orthogonal matrix of eigenvectors. This is called the spectral decomposition (see p. 94 of Gruber (2014)). Define $\mathbf{A}^{1/2} = \mathbf{P}\Delta^{1/2}\mathbf{P}'$ where $\Delta^{1/2}$ consists of the positive square roots of the elements of $\Delta$. Since $\mathbf{I} - \mathbf{A}$ is positive (semi-) definite, we have for all vectors $\mathbf{p}$,

$$\mathbf{p}'(\mathbf{I} - \mathbf{A})\mathbf{p} = \mathbf{p}'\mathbf{A}^{1/2}(\mathbf{A}^{-1} - \mathbf{I})\mathbf{A}^{1/2}\mathbf{p} \geq 0.$$

For every vector $\mathbf{q}$, there exists a $\mathbf{p}$ such that $\mathbf{q} = \mathbf{A}^{1/2}\mathbf{p}$. Thus, we have that $\mathbf{I} - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is positive semi-definite. Furthermore,

$$\mathbf{q}'(\mathbf{A}^{-1} - \mathbf{I})\mathbf{q} \geq 0.$$

Hence, $\mathbf{A}^{-1} - \mathbf{I}$ is positive (semi-) definite. Since $\mathbf{B} - \mathbf{A}$ is positive (semi-) definite

We have that $\mathbf{p}'(\mathbf{B} - \mathbf{A})\mathbf{p} = \mathbf{p}'\mathbf{B}^{1/2}(\mathbf{I} - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})\mathbf{B}^{1/2}\mathbf{p} \geq 0$ for all $\mathbf{p}$. Since for each $\mathbf{p}$ there exists a $\mathbf{q}$ where $\mathbf{q} = \mathbf{B}^{1/2}\mathbf{p}$, we have that $\mathbf{I} - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is positive (semi-) definite so that $\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2} - \mathbf{I}$ is positive (semi-) definite. Applying Lemma 2, we see that $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ is positive (semi-) definite. ■

## 4. DISTRIBUTIONS

For the sake of reference and establishing notation, some important properties of commonly used distributions will be summarized. No attempt is made at completeness or full rigor. Pertinent details that we will assume the reader is familiar with are available in many textbooks. See, for example, Hogg, Mc Kean, and Craig (2014) and Miller and Miller (2012). What follows will serve only as a reminder of these things.

### a. Multivariate Density Functions

For a set of $n$ continuous random variables $X_1, X_2, \ldots, X_n$ for which $x_1, x_2, \ldots, x_n$ represents a set of realized values, we write the cumulative density function as

$$\Pr(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = F(x_1, x_2, \ldots, x_n). \tag{5}$$

Then, the density function is

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \ldots, x_n). \tag{6}$$

A density function must satisfy these conditions:

1. $f(x_1, x_2, \ldots, x_n) \geq 0$ for $-\infty < x_i < \infty$ for all $i$;
2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n = 1$.

The marginal density function of a subset of the variables is obtained by integrating out the remaining variables in the density function. For example, if we integrate out the first $k$ variables, we obtain the marginal density of $x_k, x_{k+1,\ldots,}x_n$. Thus we have

$$g(x_{k+1}, \ldots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) dx_1 \cdots dx_k. \tag{7}$$

The conditional distribution of one subset of variables given the set of remaining variables is the ratio of the density function to the marginal density function of the remaining variables. For example, the conditional distribution of the first $k$ variables given the last $n - k$ variables is given by

$$f(x_1, \ldots, x_k | x_{k+1}, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{g(x_{k+1}, \ldots, x_n)}. \tag{8}$$

### b. Moments

The $k$th moment about zero of the $i$th variable of the expected value of the $k$th power of $x_i$ is denoted by either $\mu_{x_i}^{(k)}$ or $E(x_i^k)$. The expectation is obtained by calculating

$$\mu_{x_i}^{(k)} = \mathrm{E}(x_i^k) = \int_{-\infty}^{\infty} x_i^k g(x_i) dx_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i^k f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \tag{9}$$

When $k = 1$, the superscript $(k)$ is usually omitted and $\mu_i$ is written for $\mu_i^{(1)}$.

The covariance between the $i$th and $j$th variables for $i \neq j$ is

$$\begin{aligned} \sigma_{ij} &= \mathrm{E}(x_i - \mu_i)(x_j - \mu_j) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) g(x_i, x_j) dx_i dx_j \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n. \end{aligned} \tag{10}$$

Likewise, the variance of the $i$th variable is

$$\begin{aligned}\sigma_{ii} = \sigma_i^2 &= \mathrm{E}(x_i - \mu_i)^2 \\ &= \int_{-\infty}^{\infty} (x_i - \mu_i)^2 g(x_i) dx_i \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n.\end{aligned} \tag{11}$$

Variances and covariances between the variables in the vector $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ are given in (10) and (11). Arraying these variances and covariances as the elements of a matrix gives the *variance covariance matrix* of the $x$'s as

$$\mathrm{var}(\mathbf{x}) = \mathbf{V} = \{\sigma_i j\} \quad \text{for} \quad i, j = 1, 2, ..., n.$$

Diagonal elements of $\mathbf{V}$ are variances and off-diagonal elements are covariances.

**Notation**. The variance of a scalar random variable $x$ will be written as $\mathrm{v}(x)$. The variance-covariance matrix of a vector of random variables $\mathbf{x}$ will be denoted by $\mathrm{var}(\mathbf{x})$.

The vector of means corresponding to $\mathbf{x}'$ is

$$\mathrm{E}(\mathbf{x}') = \mu' = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \end{bmatrix}.$$

By the definition of variance and covariance,

$$\mathrm{var}(\mathbf{x}) = \mathrm{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']. \tag{12}$$

Furthermore, since the correlation is between the $i$th and $j$th variables is $\sigma_{ij}/\sigma_i\sigma_j$, the matrix of correlations is

$$\mathbf{R} = \left\{ \frac{\sigma_{ij}}{\sigma_i \sigma_j} \right\} = \mathbf{D}\{1/\sigma_i\}\mathbf{V}\mathbf{D}\{1/\sigma_j\} \quad \text{for } i, j = 1, \ldots, n \tag{13}$$

where, using (3) of Section 1.1 of Chapter 1, the $\mathbf{D}$'s are the diagonal entries with elements $1/\sigma_i$ for $i = 1, 2, \ldots, n$. The diagonal elements of $\mathbf{R}$ are all unity and $\mathbf{R}$ is symmetric. The matrix $\mathbf{R}$ is known as the correlation matrix.

The matrix $\mathbf{V}$ is non-negative definite. The variance of $\mathbf{t}'\mathbf{x}$ for any $\mathbf{t}$ is the quadratic form $\mathbf{t}'\mathbf{V}\mathbf{t} \geq 0$ that is positive by the definition of a variance unless $\mathbf{t}$ is the zero vector.

### c. Linear Transformations

When the variables $\mathbf{x}$ are transformed to variables $\mathbf{y}$ by a linear transformation $\mathbf{y} = \mathbf{T}\mathbf{x}$, it is easy to derive the moments. We have, for example,

$$\mu_y = \mathbf{T}\mu_x \text{ and } \mathrm{var}(\mathbf{y}) = \mathbf{T}\mathbf{V}\mathbf{T}'. \tag{14}$$

When making this kind of transformation with $\mathbf{T}$ non-singular, an integral involving the differentials $dx_1$, $dx_1$,…, $dx_n$ is transformed by substituting for the $x$'s in terms of the $y$'s and by replacing the differentials by $\|\mathbf{J}\|\, dy_1 dy_2 \cdots dy_n$ where $\|\mathrm{J}\|$ is the Jacobian of the $x$'s with respect to the $y$'s. The Jacobian matrix is defined as $\mathbf{J} = \left\{ {}^{\partial x_i}/_{\partial y_j} \right\}$ for $i, j = 1, 2, \ldots, n$ and $\|\mathbf{J}\|$ is the absolute value of the determinant $|\mathbf{J}|$. Since $\mathbf{x} = \mathbf{T}^{-1}\mathbf{y}$, $\mathbf{J} = \mathbf{T}^{-1}$ and $\|\mathbf{J}\| = 1/\|\mathbf{T}\|$. Hence, when the transformation from $x$ to $y$ is $\mathbf{y} = \mathbf{T}\mathbf{x}$, the product of the differentials

$$dx_1 dx_2 ... dx_n \text{ is replaced by } (dy_1 dy_2 ... dy_n)/||\mathbf{T}||. \tag{15}$$

For more information about the results in the above discussion, see, for example, Courant (1988).

The discussion above outlines the procedure for deriving the density function of $\mathbf{y} = \mathbf{T}\mathbf{x}$ from that of $\mathbf{x}$. First, substitute from $\mathbf{x} = \mathbf{T}^{-1}\mathbf{y}$ for each $x_i$ in $f(x_1, x_2, \ldots, x_n)$. Suppose that the resulting function of the $y$'s is written as $f(\mathbf{T}^{-1}\mathbf{y})$. Then, because

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n = 1$$

the transformation gives

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f\left(\mathbf{T}^{-1} y\right) (1/\,\|\mathbf{T}\|)\, dy_1 \cdots dy_n = 1.$$

Suppose that $h(y_1, y_2, \ldots, y_n)$ is the density function of the $y$'s.

Then

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, y_2, \ldots, y_n) dy_1 \cdots dy_n = 1.$$

By comparison,

$$h(y_1, y_2, \ldots, y_n) = \frac{f(\mathbf{T}^{-1}\mathbf{y})}{||\mathbf{T}||}. \tag{16}$$

**Example 4** If

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ 5 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is the transformation $\mathbf{y} = \mathbf{T}\mathbf{x}$, then $\|\mathbf{T}\| = 2$. Since

$$\mathbf{T}^{-1} = \begin{bmatrix} 2 & -1 \\ \frac{5}{2} & \frac{3}{2} \end{bmatrix}$$

$$h(y_1, y_2) = \frac{1}{2} f(2y_1 - y_2, \frac{1}{2}(5y_1 - 3y_2)).$$ □

### d. Moment and Cumulative Generating Functions

Moments and relationships between distributions are often derived using moment generating functions. In the univariate case, the moment generating function (m.g.f.) is written as a function of $t$. For a more complete discussion of the m.g.f., see, for example, Hogg, Mc Kean, and Craig (2014) or Miller and Miller (2012).

The moment generating function of a random variable $X$ is defined as the expectation of $e^{xt}$. We have that, assuming certain conditions are satisfied in order to move summation outside of the integral,

$$\begin{aligned} M_X(t) &= E(e^{xt}) \\ &= \int_{-\infty}^{\infty} e^{xt} f(x)dx = \int_{-\infty}^{\infty} \sum_{i=0}^{\infty} \frac{t^n x^n}{n!} f(x)dt = \sum_{i=0}^{\infty} \frac{\mu_X^{(i)}}{n!} t^n. \end{aligned} \tag{17}$$

The moment generating function gives a method of obtaining the central moments of a density function by differentiation. Evaluating the $k$th partial derivative of $M_X(t)$ at $t = 0$, we get that

$$m_x^{(k)} = \left. \frac{d^k M_x(t)}{dt^k} \right|_{t=0}. \tag{18}$$

Likewise, for some function of $x$, we have that

$$M_{h(x)}(t) = E(e^{th(x)}) = \int_{-\infty}^{\infty} e^{th(x)} f(x)dx. \tag{19}$$

The $k$th moment about zero of the function is

$$\mu_{h(x)}^k = \left. \frac{d^k M_{h(x)}(t)}{dt^k} \right|_{t=0}. \tag{20}$$

Similar results hold for multivariate situations. The m.g.f. of the joint distribution of $n$ variables utilizes a vector of parameters $\mathbf{t'} = \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix}$. We have that

$$M_x(t) = E(e^{t'x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t'x} f(x_1, x_2, \dots, x_n)dx_1 \cdots dx_n. \tag{21}$$

The m.g.f. of a scalar function of the elements of $\mathbf{x}$, the quadratic form $\mathbf{x'Ax}$, for example, is

$$\mathrm{M}_{\mathrm{x'Ax}}(t) = E(e^{t\mathbf{x'Ax}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t\mathbf{x'Ax}} f(x_1, x_2, \dots, x_n)dx_1 \cdots dx_n. \tag{22}$$

As well as yielding the central moments of density function, the m.g.f. has other important uses, two of which will be invoked repeatedly. First, if two random variables

have the same m.g.f., they also have the same density function. This is done under wide regularity conditions whose details are omitted here (see, for example, Mood, Graybill, and Boes (1974)). Second, two random variables are independent if their joint m.g.f. factorizes into the product of their two separate m.g.f.'s. This means that if $M_{(X_1,X_2)}(t_1, t_2) = M_{X_1}(t_1)M_{X_2}(t_2)$, then $X_1$ and $X_2$ are independent.

Another useful property of the m.g.f. is

$$M_{aX+b}(t) = e^{bt}M_X(at). \tag{23}$$

To see this, observe that

$$M_{aX+b}(t) = E(e^{t(aX+b)}) = E(e^{bt}e^{atX}) = e^{bt}M_X(at).$$

Sometimes, it is easier to obtain moments using the natural logarithm of the moment generating function. This is the culmulant generating function

$$K_X(t) = \log M_X(t). \tag{24}$$

Unless stated otherwise, log will always refer to natural logarithms with base $e$. A particularly noteworthy fact about the culmulant generating function is that its first two derivatives at zero are the mean and the variance of a density function. We have that

$$K'_X(t) = \frac{M'_X(t)}{M_X(t)}$$

so that

$$K'_X(0) = \frac{M'_X(0)}{M_X(0)} = \mu_X.$$

furthermore

$$K''_X(t) = \frac{M_X(t)M''_X(t) - (M'_X(t))^2}{(M_X(t))^2}$$

and

$$K''_X(0) = \frac{M_X(0)M''_X(0) - (M'_X(0))^2}{(M_X(0))^2} = E(X^2) - \mu^2 = \sigma^2.$$

### e. Univariate Normal

When the random variable $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$, we will write $x$ is $N(\mu, \sigma^2)$ or $x \sim N(\mu, \sigma^2)$. The density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}, \quad \text{for} -\infty < X < \infty$$

Consider the standard normal random variable $z = (x - \mu)/\sigma$. For this case, the m.g.f. of $z$ is

$$\begin{aligned} M_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{zt} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(z^2 - 2zt + t^2\right) + \frac{1}{2}t^2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} e^{\frac{1}{2}t^2} dz = e^{\frac{1}{2}t^2}. \end{aligned}$$

Now, $K_Z(t) = \frac{1}{2}t^2, K_z'(t) = t, K_z''(t) = 1$, so that for $t = 0, K_z'(0) = 0$ and $K_z''(0) = 1$. Thus, the standard normal random variable has mean 0 and variance 1. Now since $z = (x - \mu)/\sigma, x = z\sigma + \mu$.

From (23),

$$M_X(t) = e^{\mu t} M_z(\sigma t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \text{ and } \quad K_x(t) = \mu t + \frac{1}{2}\sigma^2 t^2$$

are the moment generating function and cumulant generating function, respectively of a normal random variable with mean $\mu$ and variance $\sigma^2$. Observe that the first two derivatives of $K_X(t)$ at $t = 0$ are $\mu$ and $\sigma^2$.

### f. Multivariate Normal

**(*i*) *Density Function.*** When the random variables in $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$ have a multivariate normal distribution with a vector of means $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$, we write "$\mathbf{x}$ is $N(\boldsymbol{\mu}, \mathbf{V})$" or "$\mathbf{x} \sim N(\boldsymbol{\mu},\mathbf{V})$". When $E(x_i) = \mu$ for all $i$, then $\mu = \mu\mathbf{1}$. If the $x_i$'s are mutually independent with the same variance $\sigma^2$, we write "$\mathbf{x} \sim N(\mu\mathbf{1}, \sigma^2\mathbf{I})$".

We assume that $\mathbf{V}$ is positive definite. The multivariate normal density function is then

$$f(x_1, x_2, \dots, x_n) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{V}^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{\frac{1}{2}n}|\mathbf{V}|^{\frac{1}{2}}}. \tag{25}$$

**(*ii*) *Aitken's Integral.*** We shall use Aitken's integral to show that the multivariate normal density in (25) is a bona fide probability density function. It is as follows.

For $\mathbf{A}$ being a positive definite symmetric matrix of order $n$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x}} dx_1 \cdots dx_n = (2\pi)^{\frac{1}{2}n} |\mathbf{A}|^{-\frac{1}{2}}. \tag{26}$$

We will obtain the result in (26) by means of a transformation that reduces the integral to products of single integrals of $e^{-\frac{1}{2}y_i^2}$. Notice that because $\mathbf{A}$ is positive definite, there exists a non-singular matrix $\mathbf{P}$ such that $\mathbf{P}'\mathbf{A}\mathbf{P} = \mathbf{I}_n$. Hence the determinant $|\mathbf{P}'\mathbf{A}\mathbf{P}| = |\mathbf{P}|^2|\mathbf{A}| = 1$. Thus $|\mathbf{P}| = |\mathbf{A}|^{-1/2}$. Letting $\mathbf{x} = \mathbf{P}\mathbf{y}$ gives $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{y}'\mathbf{P}'\mathbf{A}\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{y}$.

From (15),

$$\begin{aligned}
&\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x}} dx_1 \cdots dx_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}} dy_1 \cdots dy_n / ||\mathbf{P}^{-1}|| \\
&= |\mathbf{P}| \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right) dy_1 \cdots dy_n = |\mathbf{A}|^{-\frac{1}{2}} \prod_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_i^2} dy_i \right\} \\
&= (2\pi)^{\frac{1}{2}n} |\mathbf{A}|^{-\frac{1}{2}}.
\end{aligned}$$

Application of this result to (25) yields

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n = (2\pi)^{\frac{1}{2}n} \left( |\mathbf{V}^{-1}|^{-\frac{1}{2}} / \left(\sqrt{2\pi}\right)^n \right) |\mathbf{V}|^{\frac{1}{2}} = 1$$

establishing that the multivariate normal distribution is indeed a bona fide probability distribution.

***(iii) Moment Generating Function.*** The m.g.f. for the multivariate normal distribution is

$$M_{\mathrm{X}}(\mathbf{t}) = (2\boldsymbol{\pi})^{-\frac{1}{2}n} |\mathbf{V}|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}'\mathbf{x} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] dx_1 \cdots dx_n$$

Rearranging the exponent, this becomes

$$\begin{aligned}
\mathbf{M}_X(\mathbf{t}) &= (2\boldsymbol{\pi})^{-\frac{1}{2}n} |\mathbf{V}|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{V}\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{V}\mathbf{t}) + \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}] dx_1 \cdots dx_n \\
&= \frac{e^{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}}}{(2\boldsymbol{\pi})^{\frac{1}{2}n} |\mathbf{V}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{V}\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{V}\mathbf{t})] dx_1 \cdots dx_n
\end{aligned}$$

We make the transformation $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{V}\mathbf{t}$ from $\mathbf{x}$ to $\mathbf{y}$. The Jacobian is unity. The integral reduces to Aiken's integral with matrix $\mathbf{V}^{-1}$. Hence

$$M_X(\mathrm{t}) = \frac{e^{\mathrm{t}'\boldsymbol{\mu}+\frac{1}{2}\mathrm{t}'\mathbf{V}\mathrm{t}}(2\boldsymbol{\pi})^{\frac{1}{2}n}|\mathbf{V}^{-1}|^{-\frac{1}{2}}}{(2\boldsymbol{\pi})^{\frac{1}{2}n}|\mathbf{V}|^{\frac{1}{2}}} = e^{\mathrm{t}'\mu+\frac{1}{2}\mathrm{t}'\mathbf{V}\mathrm{t}}. \tag{27a}$$

The culmulant generating function

$$K_X(\mathbf{t}) = \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}. \tag{27b}$$

Finding the first and second partial derivatives of $K_{\mathrm{X}}(\mathbf{t})$ at $\mathbf{t} = 0$ gives a mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\mathbf{V}$.

**(*iv*) *Marginal Distributions.*** The definition of the marginal distribution of $x_1, x_2, \ldots, x_k$, namely the first $k$ $x$'s is in accord with (7)

$$g(x_1, x_2, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_{k+1} \cdots dx_n.$$

The m.g.f. of this distribution is by (21)

$$M_{x_1 \cdots x_k}(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1 x_1 + \cdots + t_n x_k} g(x_1, \ldots, x_k) dx_1 \cdots dx_k$$

and upon substituting for $g(x_1, x_2, \ldots, x_k)$, this becomes

$$\begin{aligned} M_{x_1 \cdots x_k}(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1 x_1 + \ldots + t_n x_k} f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\ &= \text{m.g.f}\, x_1, x_2, \ldots, x_n \text{ with } t_{k+1} = \cdots = t_n = 0 \\ &= e^{t'\mu + \frac{1}{2}t'Vt} \text{ with } t_{k+1} = \cdots = t_n = 0. \end{aligned} \tag{28}$$

To make the substitutions $t_{k+1} = \ldots t_n = 0$, we partition $\mathbf{x}$, $\boldsymbol{\mu}$, $\mathbf{V}$, and $\mathbf{t}$ by defining

$$\mathbf{x}'_1 = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix} \text{ and } \mathbf{x}'_2 = \begin{bmatrix} x_{k+1} & x_{k+2} & \cdots & x_n \end{bmatrix}.$$

Putting $\mathbf{t}_2 = \mathbf{0}$ in (28) so that $\mathbf{x}' = \begin{bmatrix} x'_1 & x'_2 \end{bmatrix}$.

Conformable with the above, we have

$$\boldsymbol{\mu}' = \begin{bmatrix} \boldsymbol{\mu}'_1 & \boldsymbol{\mu}'_2 \end{bmatrix}, \mathrm{t}' = \begin{bmatrix} \mathbf{t}'_1 & \mathbf{t}'_2 \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}'_{12} & \mathbf{V}_{22} \end{bmatrix}.$$

Putting $\mathbf{t}_2 = \mathbf{0}$ in (28) gives

$$\mathbf{M}_{x_1 \ldots x_k}(\mathbf{t}_1) = e^{\mathrm{t}'_1\mu_1 + \frac{1}{2}\mathbf{t}'_1\mathbf{V}_{11}\mathbf{t}_1}.$$

We therefore see that $g(\mathbf{x}_1)$ and $g(\mathbf{x}_2)$ are multivariate normal distributions. The marginal density functions are

$$g(\mathbf{x}_1) = g(x_1, \ldots, x_k) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}_1 - \mu_1)'\mathbf{V}_{11}(\mathbf{x}_1 - \mu_1)\right]}{(2\pi)^{\frac{1}{2}k}|\mathbf{V}_{11}|^{\frac{1}{2}}} \tag{29a}$$

and

$$g(x_2) = g(x_{k+1}, \ldots, x_n) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)'\mathbf{V}_{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right]}{(2\boldsymbol{\pi})^{\frac{1}{2}k}|\mathbf{V}_{22}|^{\frac{1}{2}}}. \tag{29b}$$

Since $\mathbf{V}$ is taken as being positive definite, so are $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$. Furthermore, in these expressions, use can be made of the partitioned form of $\mathbf{V}$ (see equation (54) of Chapter 1).

Thus, if

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}'_{12} & \mathbf{V}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}'_{12} & \mathbf{W}_{22} \end{bmatrix},$$

then $\mathbf{V}_{11}^{-1} = \mathbf{W}_{11} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{W}'_{12}$ and $\mathbf{V}_{22}^{-1} = \mathbf{W}_{22} - \mathbf{W}_{12}\mathbf{W}_{11}^{-1}\mathbf{W}'_{12}$

**(*v*) *Conditional Distributions*** Let $f(x)$ denote the density function of all $n$ $x$'s. Then equation (8) gives the conditional distribution of the first $k$ $x$'s as

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x})}{g(\mathbf{x}_2)}.$$

On substituting from (25) and (29),

$$f(\boldsymbol{x}_1|\boldsymbol{x}_2) = \frac{\exp\{-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) - (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\right]}{(2\boldsymbol{\pi})^{\frac{1}{2}k}(|\mathbf{V}|/|\mathbf{V}_{22}|)^{\frac{1}{2}}}. \tag{30}$$

In terms of the partitioned form of $\mathbf{V}$ and its inverse given above, we have

$$\mathbf{W}_{11} = (\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}'_{12})^{-1} \tag{31a}$$

and

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{W}_{11} & -\mathbf{W}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1} \\ -\mathbf{V}_{22}^{-1}\mathbf{V}'_{12}\mathbf{W}_{11} & \mathbf{V}_{22}^{-1} + \mathbf{V}_{22}^{-1}\mathbf{V}'_{12}\mathbf{W}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1} \end{bmatrix}. \tag{31b}$$

Therefore, the exponent in (30) becomes

$$\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)' & (\mathbf{x}_2-\boldsymbol{\mu}_2)'\end{bmatrix}\begin{bmatrix}\mathbf{W}_{11} & -\mathbf{W}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{12}'\mathbf{W}_{11} & \mathbf{V}_{22}^{-1}+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}'\mathbf{W}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\end{bmatrix}$$
$$\times\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\ (\mathbf{x}_2-\boldsymbol{\mu}_2)\end{bmatrix}-(\mathbf{x}_2-\boldsymbol{\mu}_2)'\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2).$$

The above expression simplifies to

$$\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)' & (\mathbf{x}_2-\boldsymbol{\mu}_2)'\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{12}'\end{bmatrix}\mathbf{W}_{11}[\mathbf{I}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}]\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\ (\mathbf{x}_2-\boldsymbol{\mu}_2)\end{bmatrix}$$
$$=[(\mathbf{x}_1-\boldsymbol{\mu}_1)-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)]'\mathbf{W}_{11}[(\mathbf{x}_1-\boldsymbol{\mu}_1)-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)]. \quad (32)$$

Furthermore, using the result for the determinant of a partitioned matrix (e.g., Searle (1966, p. 96)) from (31a)

$$|\mathbf{V}|=|\mathbf{V}_{22}||\mathbf{V}_{11}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{12}'|=|\mathbf{V}_{22}||\mathbf{W}_{11}^{-1}|.$$

Hence

$$\frac{|\mathbf{V}|}{|\mathbf{V}_{22}|}=|\mathbf{W}_{11}^{-1}|. \quad (33)$$

Substituting (32) and (33) into (30) gives

$$f(\mathbf{x}_1|\mathbf{x}_2)=\frac{\exp\{-\frac{1}{2}[(\mathbf{x}_1-\boldsymbol{\mu}_1)-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)]'\mathbf{W}_{11}[(\mathbf{x}_1-\boldsymbol{\mu}_1)-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)]'\}}{(2\pi)^{\frac{1}{2}k}|\mathbf{W}_{11}^{-1}|^{\frac{1}{2}}}. \quad (34)$$

This shows that the conditional distribution is also normal. In fact, from (34),

$$\mathbf{x}_1|\mathbf{x}_2\sim \mathrm{N}[\boldsymbol{\mu}_1+\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2),\mathbf{W}_{11}^{-1}]. \quad (35)$$

***(vi) Independence of Normal Random Variables*** A condition for the independence of sub-vectors of vectors of normal random variables is given in Theorem 2.

**Theorem 2** Suppose that the vector $\mathbf{x}'=\begin{bmatrix}x_1 & x_2 & \dots & x_n\end{bmatrix}$ is partitioned into $p$ sub-vectors $\mathbf{x}'=\begin{bmatrix}\mathbf{x}_1' & \mathbf{x}_2' & \dots & \mathbf{x}_p'\end{bmatrix}$.

Then a necessary and sufficient condition for the vectors to be mutually independent is, in the corresponding partitioning of $\mathbf{V}=\{\mathbf{V}_{ij}\}$ for $i,j=1,2,\dots,p$, that $\mathbf{V}_{ij}=\mathbf{0}$, for $i\neq j$.

*Proof.* The m.g.f. is by (27a)

$$\mathbf{M}_{\mathrm{x}}(\mathbf{t}) = \mathrm{e}^{\mathbf{t}'\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}} = \exp\left(\sum_{i=1}^{p}\mathbf{t}_i'\boldsymbol{\mu}_i + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{t}_i'\mathbf{V}_{ij}\mathbf{t}_j\right)$$

If $\mathbf{V}_{ij} = \mathbf{0}$ for $i \neq j$ this reduces to

$$\mathbf{M}_X(\mathbf{t}) = \exp\sum_{i=1}^{p}\left(\mathbf{t}_i'\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i'\mathbf{V}_{ii}\mathbf{t}_i\right) = \prod_{i=1}^{p}\exp\left(\mathbf{t}_i'\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i'\mathbf{V}_{ii}\mathbf{t}_i\right).$$

Since the m.g.f. of the joint distribution of independent sets of variables is the product of their several m.g.f.'s, we conclude that the $\mathbf{x}_i$'s are independent.

On the other hand, suppose the $p$ sub-vectors are independent each with variance covariance matrix $\mathbf{K}_{ii}$ . The m.g.f. of the joint distribution is

$$\prod_{i=1}^{p}\exp\left(\mathbf{t}_i'\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i'\mathbf{K}_{ii}\mathbf{t}_i\right) = \exp\sum_{i=1}^{p}\left(\mathbf{t}_i'\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i'\mathbf{K}_{ii}\mathbf{t}_i\right) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}\right)$$

where $\mathbf{V} = \begin{bmatrix} \mathbf{K}_{11} & 0 & \cdots & 0 \\ 0 & \mathbf{K}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{K}_{pp} \end{bmatrix}$. Hence $\mathbf{V}_{ij} = \mathbf{0}$ for $i \neq j$. ■

Note that while uncorrelated normal random variables are independent, this property does not hold for random variables in general. See, for example, Miller and Miller (2012).

### g. Central $\chi^2$, $F$, and $t$

Certain functions of independent normal random variables follow these distributions. In order to give the probability density functions, we will need the gamma distribution, which for a parameter $\alpha$ is defined by the improper integral

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx.$$

The gamma function is a generalization of the factorial function defined on the integers. We have that $\Gamma(n) = (n-1)!$ (see Exercise 4).

When $\mathbf{x} \sim N(\mathbf{0},\mathbf{I})$ then $\sum_{i=1}^{n} x_i^2$ has the central $\chi^2$-distribution with $n$ degrees of freedom. Thus when $\mathbf{x} \sim N(\mathbf{0},\mathbf{I})$ and $u = \sum_{i=1}^{n} x_i^2$ then $u \sim \chi_n{}^2$.

The density function is

$$f(u) = \frac{u^{n-1}e^{-\frac{1}{2}u}}{2^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)}, \quad u > 0. \tag{36}$$

The m.g.f. corresponding to (36) is given by (see Exercise 5)

$$M_u(t) = (1-2t)^{-\frac{1}{2}n}. \tag{37}$$

This may be obtained directly from (17) using (36) or as $M_{\mathbf{x}'\mathbf{x}}(t)$ using the density function of the standard normal distribution. The mean and the variance are $n$ and $2n$, respectively. These may be obtained by direct calculation using (36) and the properties of the gamma function or from the first two derivatives of (37) at $t = 0$.

The commonest application of the $\chi^2$-distribution is that when $\mathbf{x} \sim N(\mu\mathbf{1},\sigma^2\mathbf{I})$ then $\sum_{i=1}^{n-1}(x_i - \bar{x})^2/\sigma^2 \sim \chi^2_{n-1}$. This result can be established by making the transformation $\mathbf{y} = \mathbf{H}_0\mathrm{x}$ where $\mathbf{H}_0$ is the last $n-1$ rows of the Helmert matrix in Section 1 (see pp. 51–52 of Gruber (2014)).

The ratio of two independent random variables each having central $\chi^2$-distributions divided by the number of degrees of freedom has an $F$-distribution. Thus if

$$u_1 \sim \chi^2_{n_1} \quad \text{and} \quad u_2 \sim \chi^2_{n_2}, \quad \text{then} \quad v = \frac{u_1/n_1}{u_2/n_2} \sim F_{n_1,n_2},$$

the $F$-distribution with $n_1$ and $n_2$ degrees of freedom. The density function is

$$f(v) = \frac{\Gamma\left(\frac{1}{2}(n_1+n_2)\right) n_1^{\frac{1}{2}n_1} n_2^{\frac{1}{2}n_2} v^{\frac{1}{2}n_1-1}}{\Gamma\left(\frac{1}{2}n_1\right)\Gamma\left(\frac{1}{2}n_2\right)\left(n_2+n_1v\right)^{\frac{1}{2}n_1+\frac{1}{2}n_2}}, \quad v > 0. \tag{38}$$

The mean and the variance are

$$\mu = \frac{n_2}{n_2-2} \quad \text{and} \quad \sigma^2 = \frac{2n_1^2[1+(n_2-2)/n_1]}{(n_2-2)^2(n_2-4)}.$$

Finally, the ratio of a normally distributed random variable to the square root of one that has a chi-square distribution is the basis of the Student $t$-distribution. Thus when $x \sim N(0, 1)$ and $u \sim \chi^2_n$ independent of $x$, we have that

$$z = \frac{x}{\sqrt{\frac{u}{n}}} \sim t_n,$$

the $t$-distribution with $n$ degrees of freedom. Its density function is

$$f(x) = \frac{\Gamma\left(\frac{1}{2}(n+1)\right)}{\sqrt{n\pi}\Gamma\left(\frac{1}{2}n\right)}\left(1+\frac{z^2}{n}\right)^{-\frac{1}{2}(n+1)}, \quad \text{for} \quad -\infty < t < \infty. \tag{39}$$

A frequent application of the $t$-distribution is that if $\mathbf{x} \sim N(\mu\mathbf{1}, \sigma^2\mathbf{I})$ and if $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, then $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ has the Student $t_{n\text{-}1}$-distribution.

The relationship between $t_n$ and $F_{1,n}$ can easily be described. For $x$ that follows a standard normal distribution, consider

$$z^2 = \frac{x^2}{u/n}.$$

The random variable $x^2$ has a chi-square distribution with one degree of freedom and $u$ is chi-square with $n$ degrees of freedom. Thus $z^2 \sim F_{1,n}$ . Thus, when a variable is distributed as $t_n$ its square is distributed as $F_{1,n}$.

For the derivation of the density functions of the chi-square, $t$, and $F$-distributions, see, for example, Hogg, Mc Kean and Craig (2014).

### h. Non-central $\chi^2$

We have already seen that when $x \sim N(\mathbf{0},\mathbf{I}_n)$ distribution the distribution of $x'x = \sum x_i^2$ is what is known as the central $\chi^2$-distribution. We now consider the distribution of $\mathbf{u} = \mathbf{x}'\mathbf{x}$ when $x \sim N(\boldsymbol{\mu}, \mathrm{I})$. The sole difference is that the mean of random variable $x$ is $\mu \neq 0$. The resulting distribution of $\mathbf{u} = \mathbf{x}'\mathbf{x}$ is known as the non-central $\chi^2$.

Like the central $\chi^2$-distribution, it is the sum of the squares of independent normal random variables and involves degrees of freedom. It involves the parameter

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu} = \frac{1}{2}\sum \mu_i^2.$$

This distribution will be referred to by the symbol $\chi^{2'}(n, \lambda)$, the non-central chi-square distribution with $n$ degrees of freedom and non-centrality parameter $\lambda$. When $\mu = 0$, $\lambda = 0$, it reduces to a central chi-square distribution.

The density function of the non-central chi-square distribution $\chi^{2'}(n, \lambda)$ is

$$f(u) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\frac{u^{\frac{1}{2}n+k-1}e^{-\frac{1}{2}u}}{2^{\frac{1}{2}n+k}\Gamma(\frac{1}{2}n+k)}. \tag{40}$$

This is an infinite weighted sum of central chi-square distributions because the term $u^{\frac{1}{2}n+k}e^{-\frac{1}{2}u}/2^{\frac{1}{2}n+k}\Gamma(\frac{1}{2}n+k)$ in (40) is by (36) the density function of a central $\chi^2_{\frac{1}{2}n+k}$-

distribution. The derivation of the density function of the non-central chi-square distribution may be found in C. R. Rao (1973, p. 182).

The m.g.f. of the non-central chi-square distribution may be derived using the definition of the m.g.f. in (17) and the m.g.f. of the central chi-square distribution given by

$$\begin{aligned} M_u(t) &= e^{-\lambda} \sum_{k=0}^{\infty} (\lambda^k/k!) \left( \text{m.g.f. of} \chi^2_{\frac{1}{2}n+k} \right) \\ &= e^{-\lambda} \sum_{k=0}^{\infty} (\lambda^k/k!)(1-2t)^{-(\frac{1}{2}n+k)} \\ &= e^{-\lambda} e^{\lambda(1-2t)^{-1}} (1-2t)^{-\frac{1}{2}n} \\ &= (1-2t)^{-\frac{1}{2}n} e^{-\lambda[1-(1-2t)^{-1}]}. \end{aligned} \tag{41}$$

The culmulant generating function

$$K_u(t) = \log M_u(t) = -\frac{1}{2} n \log(1-2t) - \lambda + \frac{\lambda}{1-2t}. \tag{42}$$

The mean and variance may be obtained by finding the first two derivatives of (42) at $t = 0$. They are $n + 2\lambda$ and $2n + 8\lambda$, respectively.

As one would expect, the properties of the non-central chi-square distribution reduce to those of the central chi-square distribution when $\lambda = 0$.

The following theorem gives another important property of both the central and the non-central chi-square distribution, namely that the sum of independent chi-square random variables is also chi-square.

**Theorem 3** If for $i = 1, 2, \ldots, k$, $u_i \sim \chi^{2'}(n_i, \lambda_i)$ and independent then $\sum u_i \sim \chi^{2'}\left(\sum n_i, \sum \lambda_i\right)$.

*Proof.* We use the moment generating function and the independence of the $u_i$'s.

Observe that

$$M_{(u_1,\ldots u_k)}(\mathbf{t}) = \prod M_{u_i}(t_i) = \prod E(e^{t_i u_i}).$$

Putting $t_i = t$ for all $i$, this becomes

$$\prod M_{u_i}(t) = \prod E(e^{tu_i}) = E(e^{t\sum u_i}) = M_{\sum u_i}(t)$$

where the products and sums are over $i = 1, 2, \ldots, k$. Hence

$$\begin{aligned} M_{\sum u_i}(t) &= \prod M_{u_i}(t) = \prod (1-2t)^{-\frac{1}{2}n_i} e^{-\lambda_i[1-(1-2t)^{-1}]} \\ &= (1-2t)^{-\frac{1}{2}\sum n_i} e^{-\sum \lambda_i[1-(1-2t)^{-1}]}. \end{aligned}$$

Comparison with (41) yields $\sum u_i \sim \chi^{2'}\left(\sum n_i, \sum \lambda_i\right)$. ■

### i. Non-central *F*

Just as there is a non-central analogy of the central chi-square distribution, so also there is a non-central *F*-distribution. It is specified as follows. If $u_1$ and $u_2$ are independent random variables and if

$$u_1 \sim \chi^{2'}(n_1, \lambda) \quad \text{and} \quad u_2 \sim \chi^2_{n_2},$$

then,

$$v = \frac{u_1/n_1}{u_2/n_2} \text{ is distributed as } \mathrm{F}'(n_1, n_2, \lambda).$$

This is the non-central chi-square distribution with $n_1$ and $n_2$ degrees of freedom and non-centrality parameter $\lambda$. The density function is

$$f(v) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \frac{n_1^{\frac{1}{2}n_1+k} n_2^{\frac{1}{2}n_2} \Gamma\left[\frac{1}{2}(n_1+n_2)+k\right]}{\Gamma\left(\frac{1}{2}n_1+k\right)\Gamma\left(\frac{1}{2}n_2\right)} \frac{v^{\frac{1}{2}n_2+k-1}}{(n_2+n_1 v)^{\frac{1}{2}(n_1+n_2)+k}}.$$

When $\lambda = 0$, this reduces to (38), the density function of the central *F*-distribution. The mean and variance of the distribution are

$$\mu = \frac{n_2}{n_2-2}\left(1+\frac{2\lambda}{n_1}\right)$$

and

$$\sigma^2 = \frac{2n_2^2}{n_1^2(n_2-2)}\left[\frac{(n_1+2\lambda)^2}{(n_2-2)(n_2-4)} + \frac{n_1+4\lambda}{n_2-4}\right].$$

When $\lambda = 0$ these reduce, of course, to the mean and the variance of the central $F_{n_1,n_2}$-distribution.

### j. The Non-central *t* Distribution

If $x \sim N(\mu, 1)$ and if independently of $x$ $u \sim \chi^2_n$, then $t = x/\sqrt{u/n}$ has the non-central *t*-distribution with, $t'(n, \mu)$, with $n$ degrees of freedom and non-centrality parameter $\mu$. The density function is

$$f(t) = \frac{n^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)} \frac{e^{-\frac{1}{2}\mu^2}}{(n+t^2)^{\frac{1}{2}(n+1)}} \sum_{k=0}^{\infty} \frac{\Gamma[\frac{1}{2}(n+k+1)]\mu^k 2^{\frac{1}{2}k} t^k}{k!(n+t^2)^{\frac{1}{2}k}}.$$

Its derivation is given in C. R. Rao (1973, pp. 171–172).

## 5. DISTRIBUTION OF QUADRATIC FORMS

We discuss here the distribution of a quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ when $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$. For the most part, we will confine ourselves to the case where $\mathbf{V}$ is non-singular. In dealing with the general case of $\mathbf{x}$ being $N(\boldsymbol{\mu}, \mathbf{V})$ the results we shall obtain will also be true for important special cases such as $\mathbf{x}$ being $N(\mathbf{0},\mathbf{I})$, $N(\mu\mathbf{1}, \mathbf{I})$, or $N(\boldsymbol{\mu}, \mathbf{I})$. The main results will be presented in a series of five theorems. The first relates to culmulants of quadratic forms. The second is about the distribution of quadratic forms. The last three are about the independence properties of quadratic forms.

All of the theorems make considerable use of the trace of a matrix. The trace of a matrix is the sum of its diagonal elements. The important properties of the trace of a matrix include the following:

1. It is the sum of its eigenvalues.
2. It is equal to the rank of an idempotent matrix.
3. Products are cyclically commutative, for example,

$$\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA}) = \operatorname{tr}(\mathbf{CAB}).$$

4. For a quadratic form, we have

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \operatorname{tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{x}').$$

For more information about the trace of a matrix, see Sections 4.6 and 8.4 of Gruber (2014). The above properties of the trace are used many times in what follows without explicit reference to them. We shall assume that the reader is familiar with them.

For all of the theorems, with the exception of Theorem 4, we shall assume that $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$. The first part of Theorem 4 will hold true when the mean $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and the variance covariance matrix or dispersion $\mathrm{D}(\mathbf{x}) = \mathbf{V}$ regardless of whether $\mathbf{x}$ is normally distributed or not.

The following lemma is used to prove one result for the normal case.

**Lemma 10** For any vector $\mathbf{g}$ and any positive definite symmetric matrix $\mathbf{W}$

$$(2\pi)^{\frac{1}{2}n}|\mathbf{W}|^{\frac{1}{2}}e^{\frac{1}{2}z'Wz} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{W}^{-1}\mathbf{x} + \mathbf{g}'\mathbf{x}\right)dx_1\ldots dx_n. \quad (43)$$

*Proof.* From the integral of a multivariate normal density $\mathrm{N}(\boldsymbol{\mu}, \mathbf{W})$, we have

$$\begin{aligned}(2\pi)^{\frac{1}{2}n}|\mathbf{W}|^{\frac{1}{2}}e^{\frac{1}{2}\mathbf{z}'\mathrm{W}\mathbf{z}} &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu})dx_1\ldots dx_{\mathrm{n}}\right.\\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{W}^{-1}\mathbf{x} + \boldsymbol{\mu}'\mathbf{W}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}'\mathbf{W}^{-1}\boldsymbol{\mu}\right)dx_1\ldots dx_{\mathrm{n}}.\end{aligned}$$

On writing $\mathbf{g}'$ for $\boldsymbol{\mu}'\mathbf{W}^{-1}$ this gives (43). ∎

### a. Cumulants

Recall that the cumulant generating function was the natural logarithm of the moment generating function. It can be represented by the series

$$K_X(t) = \sum_{r=0}^{\infty} \frac{K_r}{r!} t^r \tag{44}$$

where

$$K_r = \frac{d^r}{dt^r} K_X(t)\bigg|_{t=0}$$

is the $r$th cumulant.

The first theorem about cumulants is as follows.

**Theorem 4** Without any assumptions about the distribution of $\mathbf{x}$

(i)
$$\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}; \tag{45}$$

If $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$,

(ii) the $r$th cumulant of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is
$$\mathrm{K}_r(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2^{r-1}(r-1)![\mathrm{tr}(\mathbf{A}\mathbf{V})^r + \mathbf{r}\boldsymbol{\mu}'\mathbf{A}(\mathbf{V}\mathbf{A})^{r-1}\boldsymbol{\mu}]; \tag{46}$$

and

(iii) the covariance of $\mathbf{x}$ with $\mathbf{x}'\mathbf{A}\mathbf{x}$ is
$$\mathrm{cov}(\mathbf{x}, \mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathbf{V}\mathbf{A}\boldsymbol{\mu}.$$

*Proof.* (i) With $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{x}) = \mathbf{V}$, we have

$$\mathrm{E}(\mathbf{x}\mathbf{x}') = \mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}'.$$

Hence,

$$\mathrm{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{E}\mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}') = \mathrm{tr}\mathrm{E}(\mathbf{A}\mathbf{x}\mathbf{x}') = \mathrm{tr}(\mathbf{A}\mathbf{V} + \mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}') = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.$$

Notice that the steps of the above proof only depended on the moments of $\mathbf{x}$,not on any other assumptions about its distribution.

(ii) The m.g.f. of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is

$$\boldsymbol{M}_{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}(t) = (2\pi)^{-\frac{1}{2}n}|V|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}\mathbf{x}'\mathbf{A}\mathbf{x} - \frac{1}{2}(\mathbf{x}-\mu)'\mathbf{V}^{-1}(\mathbf{x}-\mu)\right] dx_1 \ldots dx_n.$$

On rearranging the exponent, the m.g.f. becomes

$$\boldsymbol{M}_{\boldsymbol{x'Ax}}(t) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}'V^{-1}\boldsymbol{\mu}}}{(2\boldsymbol{\pi})^{\frac{1}{2}n}\,|\mathbf{V}|^{\frac{1}{2}}} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\mathbf{x}'(\mathbf{I}-2t\mathbf{AV})\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}'\mathbf{V}^{-1}\mathbf{x}\right] dx_1 \ldots dx_n. \tag{47}$$

In Lemma 10, put $\mathbf{g}' = \boldsymbol{\mu}'\mathbf{V}^{-1}$ and $\mathbf{W} = [(\mathbf{I}-2t\mathbf{AV})\mathbf{V}^{-1}]^{-1} = \mathbf{V}(\mathbf{I}-2t\mathbf{AV})^{-1}$. The right-hand side of (43) then equals the multiple integral in (47). Then (47) becomes

$$\boldsymbol{M}_{\boldsymbol{x'Ax}}(t) = e^{-\frac{1}{2}\boldsymbol{\mu}'\mathrm{V}^{-1}\boldsymbol{\mu}}\,|\mathbf{V}|^{-\frac{1}{2}}\left|\mathbf{V}(\mathbf{I}-2t\mathbf{AV})^{-1}\right|^{\frac{1}{2}} \exp\left[\frac{1}{2}\boldsymbol{\mu}'\mathbf{V}^{-1}\mathbf{V}(\mathbf{I}-2t\mathbf{AV})^{-1}\mathbf{V}^{-1}\boldsymbol{\mu}\right].$$

This simplifies to

$$\boldsymbol{M}_{\boldsymbol{x'Ax}}(t) = |(\mathbf{I}-2\mathrm{tAV})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'[\mathbf{I}-(\mathbf{I}-2\mathrm{tAV})^{-1}\mathbf{V}^{-1}\boldsymbol{\mu}]\right\}. \tag{48}$$

Recall that the cumulant generating function is the natural logarithm of the moment generating function. From (44), we have

$$\begin{aligned}\sum_{r=1}^{\infty}\frac{\boldsymbol{K}_r\mathrm{t}^r}{r!} &= \log[\boldsymbol{M}_{\boldsymbol{x'Ax}}(\mathrm{t})\\ &= -\frac{1}{2}\log|\mathbf{I}-2\mathrm{tAV}| - \frac{1}{2}\boldsymbol{\mu}'[\mathbf{I}-(\mathbf{I}-2\mathrm{tAV})^{-1}]\mathbf{V}^{-1}\boldsymbol{\mu}.\end{aligned} \tag{49}$$

The two parts of the last expression in (48) are evaluated as follows. Use "$\lambda_i$ of X" to denote the "$i$th eigenvalue of $\mathbf{X}$". Then for sufficiently small $|t|$,

$$\begin{aligned}-\frac{1}{2}\log|\mathbf{I}-2t\mathbf{AV}| &= -\frac{1}{2}\sum_{i=1}^{n}\log[\lambda_i\, of\,(\mathbf{I}-2t\mathbf{AV})]\\ &= -\frac{1}{2}\sum_{i=1}^{n}\log[1-2t(\lambda_i\, of\,\mathbf{AV})]\\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{r=0}^{\infty}-\frac{[2t(\lambda_i\, of\,\mathbf{AV})]^r}{r}\\ &= \sum_{r=1}^{\infty}\frac{2^{r-1}t^r}{r}\sum_{i=1}^{n}(\lambda_i\, of\,\mathbf{AV})^r\\ &= \sum_{r=1}^{\infty}\frac{2^{r-1}t^r}{r}(\mathrm{tr}\mathbf{AV})^r\end{aligned}$$

Using the formula for the sum of an infinite geometric series, for sufficiently small |t|,

$$\mathbf{I} - (\mathbf{I} - 2t\mathbf{AV})^{-1} = -\sum_{r=1}^{\infty} 2^r t^r (\mathbf{AV})^r.$$

Making these substitutions in (48) and equating the coefficients gives (46).

(iii) Finally, the covariance between $\mathbf{x}$ and $\mathbf{x}'\mathbf{Ax}$ is

$$\begin{aligned}
\operatorname{cov}(\mathbf{x}, \mathbf{x}'\mathbf{Ax}) &= \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})[\mathbf{x}'\mathbf{Ax} - \mathrm{E}(\mathbf{x}'\mathbf{Ax})] \\
&= \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})[\mathbf{x}'\mathbf{Ax} - \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} - \operatorname{tr}(\mathbf{AV})] \\
&= \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})[(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) + 2(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}\boldsymbol{\mu} - \operatorname{tr}(\mathbf{AV})] \\
&= \mathbf{0} + 2\mathbf{VA}\boldsymbol{\mu} - \mathbf{0} \\
&= 2\mathbf{VA}\boldsymbol{\mu}.
\end{aligned}$$

because odd moments of $(\mathbf{x} - \boldsymbol{\mu})$ are zero by the normality assumption. ■

The following corollaries are easy to establish.

**Corollary 4.1** When $\boldsymbol{\mu} = \mathbf{0}$

$$E(\mathbf{x}'\mathbf{Ax}) = \operatorname{tr}\mathbf{AV}.$$

Under normality

$$K_r(\mathbf{x}'\mathbf{Ax}) = 2^{r-1}(r - \mathbf{1})!\operatorname{tr}(\mathbf{AV})^r$$

and

$$\operatorname{cov}(\mathbf{x}, \mathbf{x}'\mathbf{Ax}) = 0.$$

These are results given by Lancaster (1954) and others.

**Corollary 4.2** An important application of the theorem is the value of its second part when $r = 2$ because then it gives the variance of $\mathbf{x}'\mathbf{Ax}$. We have that

$$v(\mathbf{x}'\mathbf{Ax}) = 2\operatorname{tr}(\mathbf{AV})^2 + 4\boldsymbol{\mu}'\mathbf{AVA}\boldsymbol{\mu}. \tag{50}$$

**Corollary 4.3** When $\mathbf{x} \sim \boldsymbol{N(0,V)}$

$$v(\mathbf{x}'\mathbf{Ax}) = 2\operatorname{tr}(\mathbf{AV})^2.$$

## b. Distributions

The following theorem gives conditions for a quadratic form of normally distributed random variables to be a non-central chi-square distribution.

**Theorem 5** When $\mathbf{x}$ is $N(\boldsymbol{\mu}, \mathbf{V})$ then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^{2'}[r(\mathbf{A}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}]$ if and only if $\mathbf{AV}$ is idempotent.

*Proof (sufficiency).* Given that $\mathbf{AV}$ is idempotent we will show that $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^{2'}[r(\mathbf{A}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}]$.

From (48), the m.g.f. of $\mathbf{x}'\mathbf{A}\mathbf{x}$ is

$$\begin{aligned} \mathrm{M}_{x'Ax}(t) &= |(\mathbf{I} - 2t\mathbf{AV})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'[\mathbf{I} - (\mathbf{I} - 2t\mathbf{AV})^{-1}]\mathbf{V}^{-1}\boldsymbol{\mu}\right\} \\ &= \prod_{i=1}^{n}(1 - 2t\lambda_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^k(\mathbf{AV})^k\right]\mathbf{V}^{-1}\boldsymbol{\mu}\right\}, \end{aligned}$$

where the $\lambda_i$ for $i = 1, 2, \ldots, n$ are the eigenvalues of $\mathbf{AV}$. We have that $\mathbf{AV}$ is idempotent so its eigenvalues are zero or one. If it has rank $r$, then $r$ of those eigenvalues are one and $n - r$ of the eigenvalues are zero. Furthermore, $(\mathbf{AV})^r = \mathbf{AV}$.

$$\begin{aligned} \mathrm{M}_{x'Ax}(t) &= \prod_{i=1}^{r}(1 - 2t)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^{\mathrm{k}}\right]\mathbf{AV}\mathbf{V}^{-1}\boldsymbol{\mu}\right\} \\ &= (1-2t)^{-\frac{1}{2}r} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'[1 - (1-2t)^{-1}]\mathbf{A}\boldsymbol{\mu}\right\} \\ &= (1-2t)^{-\frac{1}{2}r} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}[1 - (1-2t)^{-1}]\right\}. \end{aligned} \tag{51}$$

By comparison with (41), we see that $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^{2'}[r, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}]$ where $r = r(\mathbf{AV})$.

Since $\mathbf{V}$ is non-singular $r(\mathbf{AV}) = r(\mathbf{A})$. Hence, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^{2'}(r(\mathbf{A}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$. ■

*Proof (necessity).* We will now show that if $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^{2'}(r(\mathbf{A}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$ then $\mathbf{AV}$ is idempotent of rank $r$.

Since we know the distribution of $\mathbf{x}'\mathbf{A}\mathbf{x}$ its m.g.f. is given by (51) and also in the form given by (48). These two forms must be equal for all values of $\boldsymbol{\mu}$ and in particular for $\boldsymbol{\mu} = \mathbf{0}$. Substituting $\boldsymbol{\mu} = \mathbf{0}$ into (48) and (51) and equating gives

$$(1-2t)^{-\frac{1}{2}r} = |\mathbf{I} - 2t\mathbf{AV}|^{-\frac{1}{2}}.$$

Replacing $2t$ by $u$ and rearranging gives

$$(1 - u)^r = |\mathbf{I} - u\mathbf{AV}|.$$

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $\mathbf{AV}$. We then have

$$(1-u)^r = \prod_{i=1}^{n} (1-u\lambda_i).$$

Since the above equation is an identity in $u$ its right-hand side has no powers of $u$ exceeding $r$. Repeated use of this argument shows that $(n-r)$ of the eigenvalues are zero.

Thus, we can write

$$(1-u)^r = \prod_{i=1}^{r} (1-u\lambda_i).$$

Taking logarithms of both sides of the above equation and equating coefficients gives $r$ equations in $r$ unknown $\lambda$'s. All sums of the powers of the $\lambda$'s equal $r$. These have a solution $\lambda_i = 1$ for $i = 1, 2, \ldots, r$. Thus, $n - r$ latent roots are zero and $r$ of them are unity.

Therefore, by Lemma 9, $\mathbf{AV}$ is idempotent and the theorem is proved. ■

Operationally the most important part of this theorem is the sufficiency condition, namely that if $\mathbf{AV}$ is idempotent, then $\mathbf{x'Ax}$ has a non-central chi-square distribution. However, there are also occasions when the necessity condition is useful.

The theorem does, of course, have an endless variety of corollaries depending on the values of $\boldsymbol{\mu}$ and $\mathbf{V}$ and the choice of $\mathbf{A}$. For example, using it is one way to establish an important and widely used result.

**Example 4 The Distribution of a Corrected Sum of Squares of Normal Random Variables** Consider

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \mathbf{x'H}_0'\mathbf{H}_0\mathbf{x}.$$

where $\mathbf{H}_0$ is the last $n-1$ rows of the Helmert matrix discussed in Section 1 and used for $n = 4$ in Example 1. Then $\mathbf{H}_0\mathbf{H}_0' = \mathbf{I}$ and $\mathbf{H}_0'\mathbf{H}_0$ is idempotent. Hence, if $\mathbf{x} \sim N(\mu\mathbf{1}, \sigma^2\mathbf{I})$ Theorem 5 tells us that

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} \sim \chi^{2'}\left(n-1, \frac{1}{2}\mu\mathbf{1'H}_0'\mathbf{H}_0\mathbf{1}\mu\right) = \chi^2(n-1, 0)$$

because $\mathbf{1'H}_0'\mathbf{H}_0\mathbf{1} = 0$. □

Certain more direct corollaries of Theorem 5 of special interest will now be stated. These are all special cases.

**Corollary 5.1** If $\mathbf{x}$ is $N(\mathbf{0}, \mathbf{I})$ then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^2_r$ if and only if $\mathbf{A}$ is idempotent of rank $r$.

**Corollary 5.2** If $\mathbf{x}$ is $N(\mathbf{0}, \mathbf{V})$ then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^2_r$ if and only if $\mathbf{A}\mathbf{V}$ is idempotent of rank $r$.

**Corollary 5.3** If $\mathbf{x}$ is $N(\mu\mathbf{1}, \sigma^2\mathbf{I})$ then $\mathbf{x}'\mathbf{x}/\sigma^2 \text{is} \chi^{2'}(n, \frac{1}{2}\mu'\mu/\sigma^2)$.

**Corollary 5.4** If $\mathbf{x}$ is $N(\boldsymbol{\mu}, \mathbf{I})$ then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is $\chi^{2'}(r, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$ if and only if $\mathbf{A}$ is idempotent of rank $r$.

Additional special cases are easily established.

The proof of Theorem 5 is based on moment generating functions. The expressions for the cumulants of $\mathbf{x}'\mathbf{A}\mathbf{x}$ are given in (48). It shows that when $\mathbf{x}'\mathbf{A}\mathbf{x}$ has a non-central chi-square distribution, that is, when $\mathbf{A}\mathbf{V}$ is idempotent of rank $r$, the $k$th cumulant of $\mathbf{x}'\mathbf{A}\mathbf{x}$(with $\mathbf{A}$ being symmetric) is

$$K_k(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2^{k-1}(k-1)![\mathbf{r}(\mathbf{A}) + k\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}].$$

### c. Independence

Under this heading, we consider independence of:

1. a quadratic form and a linear form (Theorem 6);
2. two quadratic forms (Theorem 7);
3. sets of quadratic forms (Theorem 8a).

As was indicated above, we have a theorem for each case. Remember that two independent random variables have zero covariance but that two random variables with zero covariance might not be independent (See Miller and Miller (2012) for a counter example.). However, uncorrelated normal random variables are independent.

As promised, we first consider independence of a quadratic and a linear form.

**Theorem 6** When $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$, then $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are distributed independently if and only if $\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$.

Two things should be noted:

1. The quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ does not have to follow a non-central chi-square distribution for the theorem to hold.
2. The theorem does not involve $\mathbf{A}\mathbf{V}\mathbf{B}$, a product that does not necessarily exist.

*Proof of sufficiency.* The equality $\mathbf{BVA} = \mathbf{0}$ implies independence.

From Lemma 7, because $\mathbf{A}$ is symmetric, we have that $\mathbf{A} = \mathbf{LL}'$ for some $\mathbf{L}$ of full-column rank. Therefore, if $\mathbf{BVA} = \mathbf{0}$, $\mathbf{BVLL}' = \mathbf{0}$. Since $\mathbf{L}$ has full-column rank, $(\mathbf{L}'\mathbf{L})^{-1}$ exists (Corollary 9.1 Chapter1). Thus,

$$\mathbf{BVLL}' = \mathbf{0} \quad \text{implies} \quad \mathbf{BVLL}'\mathbf{L}(\mathbf{L}'\mathbf{L})^{-1} = \mathbf{0}, \text{ that is, } \mathbf{BVL} = \mathbf{0}.$$

Therefore,

$$\text{cov}(\mathbf{Bx}, \mathbf{x}'\mathbf{L}) = \mathbf{BVL} = \mathbf{0}.$$

Hence, because $\mathbf{x}$ is a vector of normally distributed random variables, $\mathbf{Bx}$ and $\mathbf{x}'\mathbf{L}$ are distributed independently. Consequently, $\mathbf{Bx}$ and $\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{LL}'\mathbf{x}$ are distributed independently. ■

*Proof of necessity.* The independence of $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{Bx}$ implies $\mathbf{BVA} = \mathbf{0}$.

The independence property gives $\text{cov}(\mathbf{Bx}, \mathbf{x}'\mathbf{Ax}) = \mathbf{0}$. Theorem 4(iii) gives $\text{cov}(\mathbf{Bx}, \mathbf{x}'\mathbf{Ax}) = 2\mathbf{BVA}\boldsymbol{\mu}$. Hence, $2\mathbf{BVA}\boldsymbol{\mu} = \mathbf{0}$. Since this is trnue for all $\boldsymbol{\mu}$, $\mathbf{BVA} = 0$ and the proof is complete. ■

The next theorem deals with the independence of two quadratic forms. It is similar to Theorem 6 just considered. Its proof follows the same pattern.

**Theorem 7** When $\text{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$, the quadratic forms $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{x}'\mathbf{Bx}$ are distributed independently if and only if $\mathbf{AVB} = \mathbf{0}$ (or equivalently $\mathbf{BVA} = \mathbf{0}$).

Note that the form of the distributions of $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{x}'\mathbf{Bx}$ is not specified in this theorem. It applies regardless of the distribution that these quadratic forms follow, provided only that $\mathbf{x}$ is a vector of normal random variables.

*Proof.* The conditions $\mathbf{AVB} = \mathbf{0}$ and $\mathbf{BVA} = \mathbf{0}$ are equivalent because $\mathbf{A}$, $\mathbf{V}$, and $\mathbf{B}$ are symmetric. Each condition therefore implies the other.
*Sufficiency:* The condition $\mathbf{AVB} = \mathbf{0}$ implies independence.

By Lemma 7, we can write $\mathbf{A} = \mathbf{LL}'$ and $\mathbf{B} = \mathbf{MM}'$ where each of $\mathbf{L}$ and $\mathbf{M}$ has full-column rank. Therefore, if $\mathbf{AVB} = \mathbf{0}$, $\mathbf{L}'\mathbf{LL}'\mathbf{VMM}'\mathbf{M} = \mathbf{0}$. Since $(\mathbf{L}'\mathbf{L})^{-1}$ and $(\mathbf{M}'\mathbf{M})^{-1}$ exist, $\mathbf{L}'\mathbf{VM} = \mathbf{0}$. Therefore, $\text{cov}(\mathbf{L}'\mathbf{x}, \mathbf{x}'\mathbf{M}) = \mathbf{L}'\mathbf{VM} = \mathbf{0}$. Hence, because $\mathbf{x}$ is a vector of normally distributed random variables, $\mathbf{L}'\mathbf{x}$ and $\mathbf{x}'\mathbf{M}$ are distributed independently. Consequently, $\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{LL}'\mathbf{x}$ and $\mathbf{x}'\mathbf{Bx} = \mathbf{x}'\mathbf{MM}'\mathbf{x}$ are distributed independently.[2]

*Necessity:* Independence implies $\mathbf{AVB} = \mathbf{0}$.

[2] S. R. Searle is grateful for discussions with D. L. Solomon and N. S. Urquhart about the proofs of sufficiency in Theorems 6 and 7. Proofs can also be established, very tediously, using moment generating functions.

When $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{x}'\mathbf{B}\mathbf{x}$ are distributed independently, $\text{cov}(\mathbf{x}'\mathbf{A}\mathbf{x}, \mathbf{x}'\mathbf{B}\mathbf{x}) = \mathbf{0}$ so that

$$\text{v}(\mathbf{x}'(\mathbf{A}+\mathbf{B})\mathbf{x}) = \text{v}(\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{x}'\mathbf{B}\mathbf{x}) = \text{v}(\mathbf{x}'\mathbf{A}\mathbf{x}) + \text{v}(\mathbf{x}'\mathbf{B}\mathbf{x}) \tag{52}$$

Applying equation (50) to the first and last expression in equation (52) after some simplification we obtain (Exercise 23)

$$\text{tr}(\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{B}) + 2\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{B}\boldsymbol{\mu} = 0. \tag{53}$$

Equation (53) is true for all $\boldsymbol{\mu}$, including $\boldsymbol{\mu} = \mathbf{0}$, so that $\text{tr}(\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{B}) = 0$. On substituting back in (53) we have $2\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{B}\boldsymbol{\mu} = 0$. This in turn is true for all $\boldsymbol{\mu}$ and so $\mathbf{A}\mathbf{V}\mathbf{B} = \mathbf{0}$. Thus, the theorem is proved. ■

Before turning to the final theorem concerning independence, Theorem 8a, recall that Theorems 6 and 7 are concerned with independence properties only, and apply whether or not the quadratic forms have chi-square distributions. This is not the case with Theorem 8a. It relates to the independence of quadratic forms in a sum of quadratics and is concerned with conditions under which such forms have a non-central chi-square distribution. As such it involves idempotent matrices. Corollary 8.1, a special case of Theorem 8a, is Cochran's (1936) theorem, a very important result for the analysis of variance. The theorem follows. Its statement is lengthy.

**Theorem 8a** Let the following be given:

$\mathbf{X}$, order $n \times 1$, distributed as $N(\boldsymbol{\mu}, \mathbf{V})$;

$$A_i, n \times n, \text{ symmetric, of rank } k_i, \text{ for } i = 1, 2, \ldots, p;$$

and

$$\mathbf{A} = \sum_{i=1}^{p} \mathbf{A}_i \text{ which is symmetric with rank } k.$$

Then,

$$\mathbf{x}'\mathbf{A}_i\mathbf{x} \text{ is } \chi^{2\prime}\left(k_i, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}\right),$$

and the $\mathbf{x}'\mathbf{A}_i\mathbf{x}$ are pairwise independent and

$$\mathbf{x}'\mathbf{A}\mathbf{x} \text{ is } \chi^{2\prime}\left(k, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right)$$

if and only if

I: any 2 of (a) $\mathbf{A}_i\mathbf{V}$ idempotent for all $i$
(b) $\mathbf{A}_i\mathbf{V}\mathbf{A}_j = \mathbf{0}$ for all $i < j$
(c) $\mathbf{A}\mathbf{V}$ idempotent
are true; or

II: (c) is true and (d) $k = \sum_{i=1}^{n} k_i$; or

III: (c) is true and (e) $\mathbf{A}_1\mathbf{V}, \ldots, \mathbf{A}_{(p-1)}\mathbf{V}$ are idempotent and $\mathbf{A}_p\mathbf{V}$ is non-negative definite.

The proof of this theorem in statistics rests upon a theorem in matrices which in turn depends on a lemma. The matrix Theorem 8a given below as Theorem 8b is an extension of Graybill (1961, Theorems 1.68 and 1.69). The proof given by Graybill and Marsaglia (1957) is lengthy. The proof that will be given here follows the much shorter proof of Banerjee (1964) as improved by Loynes (1966) based upon a lemma. Accordingly, we first state and prove the lemma given by Loynes.

**Loynes' Lemma.** If $\mathbf{B}$ is symmetric and idempotent, if $\mathbf{Q}$ is symmetric and non-negative definite, and if $\mathbf{I} - \mathbf{B} - \mathbf{Q}$ is non-negative definite, then $\mathbf{BQ} = \mathbf{QB} = \mathbf{0}$.

*Proof of Loyne's Lemma.* Let $\mathbf{x}$ be any vector and let $\mathbf{y} = \mathbf{Bx}$. Then

$$\mathbf{y}'\mathbf{B}\mathbf{y} = \mathbf{y}'\mathbf{B}^2\mathbf{x} = \mathbf{y}'\mathbf{B}\mathbf{x} = \mathbf{y}'\mathbf{y}.$$

Then

$$\mathbf{y}'(\mathbf{I} - \mathbf{B} - \mathbf{Q})\mathbf{y} = -\mathbf{y}'\mathbf{Q}\mathbf{y}.$$

Furthermore, since $\mathbf{I} - \mathbf{B} - \mathbf{Q}$ is non-negative definite,

$$\mathbf{y}'(\mathbf{I} - \mathbf{B} - \mathbf{Q})\mathbf{y} \geq 0.$$

Hence $-\mathbf{y}'\mathbf{Q}\mathbf{y} \geq 0$ and so because $\mathbf{Q}$ is non-negative definite, we have that $\mathbf{y}'\mathbf{Q}\mathbf{y} = 0$. In addition, since $\mathbf{Q}$ is symmetric, $\mathbf{Q} = \mathbf{L}'\mathbf{L}$ for some $\mathbf{L}$ and therefore $\mathbf{y}'\mathbf{Q}\mathbf{y} = \mathbf{y}'\mathbf{L}'\mathbf{L}\mathbf{y} = 0$ implies $\mathbf{L}\mathbf{y} = 0$ and hence $\mathbf{L}'\mathbf{L}\mathbf{y} = 0$; that is, $\mathbf{Q}\mathbf{y} = \mathbf{QB}\mathbf{x} = 0$. Since this is true for any $\mathbf{x}$, $\mathbf{QB} = \mathbf{0}$ and so

$$(\mathbf{QB})' = \mathbf{B}'\mathbf{Q}' = \mathbf{BQ} = 0.$$

■

The Lemma is thus proved.

We will now state and prove the matrix theorem.

**Theorem 8b** Let the following be given:

$$\mathbf{X}_i, n \times n, \text{ symmetric, rank } k_i, i = 1, 2, \ldots p.$$
$$\mathbf{X} = \sum_{i=1}^{p} X_i, \text{ which is symmetric, with rank } k.$$

Then of the conditions

(a) $\mathbf{X}_i$, idempotent for all $i$,
(b) $\mathbf{X}_i\mathbf{X}_j = \mathbf{0}$ for $i \neq j$,
(c) $\mathbf{X}$ idempotent
(d) $k = \sum_{i=1}^{p} k_i$

it is true that

I. any two of (a), (b) and (c) imply (a), (b), (c), and (d);
II. (c) and (d) imply (a) and (b); and
III. (c) $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_{p-1}$ being idempotent with $\mathbf{X}_p$ being non-negative definite, imply that $\mathbf{X}_p$ is idempotent also and hence (a); and therefore (b) and (d).

Theorems 8a and 8b are analogous. Once 8b is proved, the proof of Theorem 8a is relatively brief. The part played by Theorem 8b is that it shows that in situations in which any one of section I, II, and III of Theorem 8a holds true, all of the conditions (a), (b), and (c) of section I will hold. The consequences of Theorem 5, the independence of quadratics and their chi-square distributions, then arise directly from Theorems 5 and 7.

*Proof of Theorem 8b.* We first prove section I, doing it in four parts.

I(i): Given (c), $\mathbf{I} - \mathbf{X}$ is idempotent and hence non-negative definite. Furthermore $\mathbf{X} - \mathbf{X}_i - \mathbf{X}_j = \sum_{r \neq i \neq j} \mathbf{X}_r$ is non-negative definite. Then by Loynes' Lemma, $\mathbf{X}_i\mathbf{X}_j = \mathbf{0}$ which is (b). Hence (a) and (c) imply (b).

I(ii): Let $\lambda$ be an eigenvalue and $\mathbf{u}$ the corresponding eigenvector of $\mathbf{X}_1$. Then $\mathbf{X}_1\mathbf{u} = \boldsymbol{\lambda}\mathbf{u}$ and for $\lambda \neq \mathbf{0}$, $\mathbf{u} = \mathbf{X}_1\mathbf{u}/\lambda$. Hence given (b) $\mathbf{X}_1\mathbf{u} = \mathbf{X}_i\mathbf{X}_1\mathbf{u}/\lambda = \mathbf{0}$. Therefore $\mathbf{X}\mathbf{u} = \mathbf{X}_1\mathbf{u} = \lambda\mathbf{u}$ and $\lambda$ is an eigenvalue of $\mathbf{X}$. However, given (c), $\mathbf{X}$ is idempotent and hence $\lambda = 0$ or 1. In a like manner, it can be shown that the other $\mathbf{X}_i$'s are idempotent establishing (c). Hence (b) and (c) imply (a).

I(iii): Given (b) and (a) $\mathbf{X}^2 = \sum \mathbf{X}_i^2 = \sum \mathbf{X}_i = \mathbf{X}$ which is (c).

I(iv): Given (c), $r(\mathbf{X}) = \text{tr}(\mathbf{X})$. Then

$$k = r(\mathbf{X}) = \text{tr}(\mathbf{X}) = \text{tr}(\sum \mathbf{X}) = \sum \text{tr}(\mathbf{X}_i).$$

From (a), $\sum \text{tr}(\mathbf{X}_i) = \sum k_i$. Hence $k = \sum k_i$ which is (d). Thus (a) and (c) imply (d).

II: The proof of this section follows Loynes (1966). Given (c), $\mathbf{I} - \mathbf{X}$ is idempotent. Therefore $\mathbf{X} - \mathbf{I}$ has rank $n - k$. This means that $\mathbf{X} - \mathbf{I}$ has $n - k$ linearly independent

(LIN) rows. Therefore

$$\begin{aligned} &\text{in } (\mathbf{X}-\mathbf{I})\mathbf{x}=\mathbf{0} \text{ there are } n-k \text{ LIN equations;}\\ \text{and}\quad &\text{in } \mathbf{X}_2\mathbf{x}=\mathbf{0} \text{ there are } k_2 \text{ LIN equations;}\\ &\qquad\vdots \qquad\qquad \vdots\\ \text{and}\quad &\text{in } \mathbf{X}_p\mathbf{x}=\mathbf{0} \text{ there are } k_p \text{ LIN equations.} \end{aligned}$$

However, these LIN sets of equations are not all mutually LIN. For example, the $k_2$ LIN equations in $\mathbf{X}_2\mathbf{x}=\mathbf{0}$ may not be LIN of the $k_p$ LIN equations in $\boldsymbol{X}_p x=\mathbf{0}$. Therefore in

$$\begin{bmatrix} \mathbf{X}-\mathbf{I} \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_{\mathrm{p}} \end{bmatrix}\mathbf{x}=\mathbf{0}$$

the maximum number of LIN equations is given (d),

$$n-k+k_2+\cdots+k_p=n-k_1;$$

and the equations reduce to $\mathbf{X}_1\mathbf{x}=\mathbf{x}$. Thus, the minimum number of LIN solutions to $\mathbf{X}_1\mathbf{x}=\mathbf{x}$ is $n-(n-k_1)=k_1$. That means that for at least $k_1$ LIN vectors $\mathbf{x}$, $\mathbf{X}_1\,\mathbf{x}=\mathbf{x}=1\mathbf{x}$.

Hence, 1 is an eigenvalue of $\mathbf{X}_1$ with multiplicity at least $k_1$. However, $r(\mathbf{X}_1)=k_1$. Then $\mathbf{X}_1$ has only $k_1$ non-zero eigenvalues and thus by Lemma 8 is idempotent. In a like manner, it can be shown that the other $\mathbf{X}_i$'s are idempotent and thus is (a) established. Thus, (c) and (d) imply (a) and hence by I(i), (b). Thus, II is proved.

III: Given (c), $\mathbf{X}$ is non-negative definite and then so is $\mathbf{I}-\mathbf{X}$. With $\mathbf{X}_1, \ldots, \mathbf{X}_{p\text{-}1}$ being idempotent and hence positive semi-definite, and $\mathbf{X}_p$ also then

$$\sum_{r\neq i\neq j}^{p}\mathbf{X}_r=\mathbf{X}-\mathbf{X}_i-\mathbf{X}_j \text{ is non-negative definite.}$$

Therefore,

$$\mathbf{I}-\mathbf{X}+\mathbf{X}-\mathbf{X}_i-\mathbf{X}_j=\mathbf{I}-\mathbf{X}_i-\mathbf{X}_j \text{ is non-negative definite.}$$

Then by Loynes' Lemma, $\mathbf{X}_i\mathbf{X}_j=\mathbf{0}$ so (b) is true. Therefore, (a) and (d) are implied also, and both this section and the whole theorem is proved.

We now have to show how Theorem 8b leads to proving Theorem 8a. ■

*Proof of Theorem 8a.* Since $\mathbf{V}$ is symmetric and positive definite, by Lemma 4,$\mathbf{V}=\mathbf{T}'\mathbf{T}$ for some non-singular $\mathbf{T}$. Then since $\mathbf{A}_i$ is symmetric, so is $\mathbf{T}\mathbf{A}_i\mathbf{T}'$ and $r(\mathbf{A}_i)=r(\mathbf{T}\mathbf{A}_i\mathbf{T}')$.

Furthermore, $\mathbf{A}_i\mathbf{V}$ is idempotent if and only if $\mathbf{T}\mathbf{A}_i\mathbf{T}'$ is. Also $\mathbf{A}_i\mathbf{V}\mathbf{A}_j=\mathbf{0}$ if and only if $\mathbf{T}\mathbf{A}_i\mathbf{T}'\mathbf{T}\mathbf{A}_j\mathbf{T}'=\mathbf{0}$. Hence Theorem 8b holds true using $\mathbf{T}\mathbf{A}_i\mathbf{T}'$ in place of $\mathbf{X}_i$.

and $\mathbf{TAT}'$ in place of $\mathbf{X}$. Then sections I, II, and III of Theorem 8b applied to $\mathbf{TA}_i\mathbf{T}'$ and $\mathbf{TAT}'$ show that when sections I, II, and III of Theorem 8a exist, conditions (a), (b), and (c) always exist. However, by Theorem 5, $\mathbf{x}'\mathbf{A}_i\mathbf{x}$ is $\chi^{2'}(k_i, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu})$ if and only if (a) is true. Also, $\mathbf{x}'\mathbf{Ax}$ is $\chi^{2'}(k, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$ if and only if (c) is true. By Theorem 7 $\mathbf{x}'\mathbf{A}_i\mathbf{x}$ and $\mathbf{x}'\mathbf{A}_j\mathbf{x}$ are independent if and only if condition (b) is true. And so Theorem 8a is proved. ■

The following corollary is fundamental for analysis of variance. It is the well-known theorem first proved by Cochran in 1934.

**Corollary 8.1** (Cochran's Theorem). When $\mathbf{x}$ is $N(\mathbf{0}, \mathbf{I}_n)$ and $\mathbf{A}_i$ is symmetric of rank $n$ for $i = 1, 2, \ldots, p$ with $\sum_{i=1}^{p} \mathbf{A}_i = n$, the $\mathbf{x}'\mathbf{A}_i\mathbf{x}$ are distributed independently as $\chi^2_{r_i}$ if and only if $\sum_{i=1}^{n} r_i = n$.

*Proof.* Put $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{V} = \mathbf{I}_n = \mathbf{A}$ in Theorem 8a. ■

**Example 5 An Illustration of the Ideas in Theorem 8a and 8b** Let

$$\mathbf{A} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix},$$

$$\mathbf{A}_2 = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}, \text{ and } \mathbf{A}_3 = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

It is easily shown that $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3, \mathbf{A}_1\mathbf{A}_2 = \mathbf{0}, \mathbf{A}_1\mathbf{A}_3 = \mathbf{0}, \mathbf{A}_2\mathbf{A}_3 = \mathbf{0}$, $\mathbf{A}$, $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ are idempotent matrices and $\mathbf{A}$ has rank 3, $\mathbf{A}_1, \mathbf{A}_2$ and $\mathbf{A}_3$ have rank 1. Thus, if $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I}_4)$ $\mathbf{y}'\mathbf{Ay} \sim \chi^{2'}$ $(3, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$. Furthermore, $\mathbf{y}'\mathbf{A}_i\mathbf{y} \sim \chi^{2'}(1, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}), i = 1, 2, 3$ and are distributed independently. □

For the case where $\mathbf{V}$ is singular, some similar theorems hold true. See Searle (1971) and Anderson (2003).

## 6. BILINEAR FORMS

Knowing the distributional properties of quadratic forms enables us to discuss the properties of bilinear forms. We consider the general bilinear form $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are of order $n_1$ and $n_2$ and distributed as $N(\mu_1, \mathbf{C}_{11})$ and $N(\mu_2, \mathbf{C}_{22})$, respectively. The matrix of covariances between $\mathbf{x}_1$ and $\mathbf{x}_2$ is $\mathbf{C}_{12}$ of order $n_1 \times n_2$; that is,

$$\mathrm{E}(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_2 - \boldsymbol{\mu}_2)' = \mathbf{C}_{12}.$$

Properties of the bilinear form are readily derived from those of quadratic forms because $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$ can be expressed as a quadratic form:

$$\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 = \frac{1}{2}\begin{bmatrix}\mathbf{x}_1' & \mathbf{x}_2'\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} \text{ with } \mathbf{A}_{21} = \mathbf{A}_{12}'$$

Hence,

$$\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 = \frac{1}{2}\mathbf{y}'\mathbf{B}\mathbf{y}.$$

where

$$\mathbf{B} = \mathbf{B}' = \begin{bmatrix}\mathbf{0} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{0}\end{bmatrix} \text{ with } \mathbf{A}_{21} = \mathbf{A}_{12}'$$

and

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V}) \text{ with } \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \mathbf{V} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}, \text{ where } \mathbf{C}_{21} = \mathbf{C}_{12}'.$$

These properties of $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$ are equivalent to $\frac{1}{2}(\mathbf{y}'\mathbf{B}\mathbf{y})$ which for some purposes is better viewed as $\mathbf{y}'(\frac{1}{2}\mathbf{B})\mathbf{y}$.

Similar to Theorem 4, we have that the mean of $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$, whether the distribution of the $\mathbf{x}$'s are normal or not, is

$$E(\mathbf{x}'_1\mathbf{A}_{12}\mathbf{x}_2) = \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{21}) + \boldsymbol{\mu}_1'\mathbf{A}_{12}\boldsymbol{\mu}_2. \tag{54}$$

This may be proved either in the same manner as part (i) of Theorem 4 or by applying Theorem 4 to the quadratic form $\mathbf{y}'(\frac{1}{2}\mathbf{B})\mathbf{y}$. From part (ii) of Theorem 4, we have the $r$th cumulant of $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$ as

$$\mathbf{K}_r(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2) = \frac{1}{2}(r-1)![\mathrm{tr}(\mathbf{B}\mathbf{V})^r + r\boldsymbol{\mu}'\mathbf{B}(\mathbf{V}\mathbf{B})^{r-1}\boldsymbol{\mu}]. \tag{55}$$

From Theorem 4, $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 \sim \chi^{2'}[\mathrm{r}(\mathbf{B}), \frac{1}{4}\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\mu}]$ if and only if $\frac{1}{2}\mathbf{BV}$ is idempotent. With

$$\mathbf{BV} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{C}_{21} & \mathbf{A}_{12}\mathbf{C}_{22} \\ \mathbf{A}_{21}\mathbf{C}_{11} & \mathbf{A}_{21}\mathbf{C}_{12} \end{bmatrix},$$

notice in general, idempotency of $\frac{1}{2}\mathbf{BV}$ does not imply(nor is it implied by) idempotency of $\mathbf{BV}$. In substituting $\mathbf{BV}$ into (55), use is made of $(\mathbf{A}_{21})' = \mathbf{A}_{12}$ and $(\mathbf{C}_{21})' = \mathbf{C}_{12}$ and also the cyclic commutativity of matrix products under the trace operation. Thus,

$$\mathrm{tr}(\mathbf{A}_{21}\mathbf{C}_{12}) = \mathrm{tr}(\mathbf{C}_{12}\mathbf{A}_{21}) = \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{21})' = \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{21}). \tag{56}$$

A special case of (55) when $r = 2$:

$$\mathrm{v}(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_1) = \frac{1}{2}\left[\mathrm{tr}(\mathbf{BV})^2 + 2\boldsymbol{\mu}'\mathbf{BVB}\boldsymbol{\mu}\right].$$

Substituting for $\mathbf{BV}$ and $\boldsymbol{\mu}$ and using (56) reduces this to

$$\begin{aligned} v(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2) &= \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{21})^2 + \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{22}\mathbf{A}_{21}\mathbf{C}_{11}) + \boldsymbol{\mu}_1'\mathbf{A}_{12}\mathbf{C}_{22}\mathbf{A}_{21}\boldsymbol{\mu}_1 \\ &\quad + \boldsymbol{\mu}_2'\mathbf{A}_{21}\mathbf{C}_{11}\mathbf{A}_{12}\boldsymbol{\mu}_2 + 2\boldsymbol{\mu}_1'\mathbf{A}_{12}\mathbf{C}_{21}\mathbf{A}_{12}\boldsymbol{\mu}_2. \end{aligned} \tag{57}$$

We now derive the covariance between two bilinear forms $\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2$ and $\mathbf{x}_3'\mathbf{A}_{34}\mathbf{x}_4$ based on procedures developed by Evans (1969). Let $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, and $\mathbf{x_4}$ have orders $n_1, n_2, n_3$, and $n_4$, respectively and be normally distributed with respective means $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ and covariance matrices $\mathbf{C}_{ij}$ of order $n_i \times n_j$ for $i, j = 1, 2, 3, 4$:

$$\mathbf{C}_{ij} = E(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_j)'. \tag{58}$$

Also define

$$\mathbf{x}' = \begin{bmatrix} \mathbf{x}_1' & \mathbf{x}_2' & \mathbf{x}_3' & \mathbf{x}_4' \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu}' = \begin{bmatrix} \boldsymbol{\mu}_1' & \boldsymbol{\mu}_2' & \boldsymbol{\mu}_3' & \boldsymbol{\mu}_4' \end{bmatrix} \tag{59}$$

with

$$\mathbf{C} = \{\mathbf{C}_{ij}\} \quad \text{for} \quad i, j = 1, 2, 3, 4 \tag{60}$$

for $\mathbf{C}_{ij}$ of (58). Thus, $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{C})$. Then with $\mathbf{A}_{43} = \mathbf{A}_{34}$

$$\mathbf{W} = \frac{1}{2}\begin{bmatrix} \mathbf{0} & \mathbf{A}_{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{34} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{0} \end{bmatrix}, \tag{61}$$

$$\mathbf{x}'\mathbf{W}\mathbf{x} = \mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2 + \mathbf{x}_3'\mathbf{A}_{34}\mathbf{x}_4$$

so that

$$2\,\mathrm{cov}(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}'_3\mathbf{A}_{34}\mathbf{x}_4) = \mathrm{v}(\mathbf{x}'\mathbf{W}\mathbf{x}) - \mathrm{v}(\mathbf{x}'_1\mathbf{A}_{12}\mathbf{x}_2) - \mathrm{v}(\mathbf{x}'_3\mathbf{A}_{34}\mathbf{x}_4). \tag{62}$$

Corollary 4.2 applied to the first term of (62) gives

$$v(\mathbf{x}'\mathbf{W}\mathbf{x}) = 2\mathrm{tr}(\mathbf{W}\mathbf{C})^2 + 4\boldsymbol{\mu}'\mathbf{W}\mathbf{C}\mathbf{W}\boldsymbol{\mu},$$

for $\boldsymbol{\mu}$, $\mathbf{C}$, and $\mathbf{W}$ of (60), (61), and (62), respectively. Using (57) for $v(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2)$ and its analogue for $\mathrm{v}(\mathbf{x}_3'\mathbf{A}_{34}\mathbf{x}_4)$, we find that (62) reduces, after repetitive use of the properties illustrated in (56), to

$$\begin{aligned}\mathrm{cov}(\mathbf{x}_1'\mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_3'A_{34}\mathbf{x}_4) &= \mathrm{tr}(\mathbf{A}_{12}\mathbf{C}_{23}\mathbf{A}_{34}\mathbf{C}_{41} + \mathbf{A}_{12}\mathbf{C}_{24}\mathbf{A}_{43}\mathbf{C}_{31})\\ &+ \boldsymbol{\mu}_1'\mathbf{A}_{12}\mathbf{C}_{23}\mathbf{A}_{34}\boldsymbol{\mu}_4 + \boldsymbol{\mu}_1'\mathbf{A}_{12}\mathbf{C}_{24}\mathbf{A}_{43}\boldsymbol{\mu}_3\\ &+ \boldsymbol{\mu}_2'\mathbf{A}_{21}\mathbf{C}_{13}\mathbf{A}_{34}\boldsymbol{\mu}_4 + \boldsymbol{\mu}_2'\mathbf{A}_{21}\mathbf{C}_{14}\mathbf{A}_{43}\boldsymbol{\mu}_3.\end{aligned} \tag{63}$$

This result does, of course, yield results obtained earlier when used for special cases. For example, to obtain $\mathrm{var}(\mathbf{x}'\mathbf{A}\mathbf{x})$ put all $\mathbf{A}_{ij}$'s equal to $\mathbf{A}$, all $\mathbf{C}_{ij}$'s equal to $\mathbf{V}$ and all $\mu_i$'s equal to $\mu$. Then if $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$, we have that

$$v(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathrm{tr}(\mathbf{A}\mathbf{V})^2 + 4\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{A}\boldsymbol{\mu} \tag{64}$$

as in (50). Also to obtain the covariance of two quadratic forms in the same variables say $\mathbf{x}'\mathbf{P}\mathbf{x}$ and $\mathbf{x}'\mathbf{Q}\mathbf{x}$, put all of the $\mu$'s in (63) equal to $\mu$, all the $\mathbf{C}$'s equal to $\mathbf{V}$, put $\mathbf{A}_{12} = \mathbf{A}_{21} = \mathbf{P}$ and $\mathbf{A}_{34} = \mathbf{A}_{43} = \mathbf{Q}$ to give

$$\mathrm{cov}(\mathbf{x}'\mathbf{P}\mathbf{x}, \mathbf{x}'\mathbf{Q}\mathbf{x}) = 2\mathrm{tr}(\mathbf{P}\mathbf{V}\mathbf{Q}\mathbf{V}) + 4\boldsymbol{\mu}'\mathbf{P}\mathbf{V}\mathbf{Q}\boldsymbol{\mu}. \tag{65}$$

## 7. EXERCISES

**1** Suppose the data for five observations in a row-by column analysis are as follows

| Row | Column | |
|---|---|---|
| | 1 | 2 |
| 1 | 6 | 4 |
| 2 | 6, 42 | 12 |

The analogy for unbalanced data of the interaction sum of squares is

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{y_{ij.}^2}{n_{ij}} - \sum_{i=1}^{r}\frac{y_{i..}^2}{n_{i.}} - \sum_{j=1}^{c}\frac{y_{.j.}^2}{n_{.j}} + \frac{y_{..}^2}{n_{..}}.$$

Use the above data to show that this expression is not a positive definite form. Why then can it not be described as a sum of squares?

**2** Let $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Show by direct computation that

**(a)** $\mathbf{B} - \mathbf{A}$ is non-negative definite.

**(b)** $\mathbf{A}^{-1} - \mathbf{B}^{-1}$ is non-negative definite.

**(c)** $\mathbf{B}^2 - \mathbf{A}^2$ is not non-negative definite.

**3** Let

$$\mathbf{B}_1 = \frac{1}{6}\begin{bmatrix} 2 & 2 & -1 & -1 & -1 & -1 \\ 2 & 2 & -1 & -1 & -1 & -1 \\ -1 & -1 & 2 & 2 & -1 & -1 \\ -1 & -1 & 2 & 2 & -1 & -1 \\ -1 & -1 & -1 & -1 & 2 & 2 \\ -1 & -1 & -1 & -1 & 2 & 2 \end{bmatrix}, \mathbf{B}_2 = \frac{1}{6}\begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

$$\mathbf{B}_3 = \frac{1}{6}\begin{bmatrix} 2 & -2 & -1 & 1 & -1 & 1 \\ -2 & 2 & 1 & -1 & 1 & -1 \\ -1 & 1 & 2 & -2 & -1 & 1 \\ 1 & -1 & -2 & 2 & 1 & -1 \\ -1 & 1 & -1 & 1 & 2 & -2 \\ 1 & -1 & 1 & -1 & -2 & 2 \end{bmatrix}, \mathbf{B}_4 = \frac{1}{6}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Establish that the above matrices satisfy the hypothesis of Cochran's Theorem. If $\mathbf{x} \sim N(0, \mathbf{I}_6)$ what are the distributions of $\mathbf{X}'\mathbf{A}_i\mathbf{x}, \mathbf{i} = 1, 2, 3, 4$. Are they independent?

**4** Show that

**(a)** $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

**(b)** $\Gamma(n) = (n - 1)!$

**5** **(a)** Derive the moment generating function of the $\chi^2_{\mathrm{n}}$-distribution two ways:

**(i)** from its density function;

**(ii)** from the density function of the $N(0, 1)$-distribution.

**(b)** Use your result to obtain the mean and variance of the $\chi^2_n$-distribution.

**6** **(a)** Using the result of Exercise 5a derive the cumulant generating function of the $\chi^2_n$-distribution.

**(b)** Use the result of (a) to obtain the mean and variance of the $\chi^2_n$-distribution again.

**7** Find the first two moments of $1/u$ when

**(a)** $u \sim \chi^2_n$

**(b)** $u \sim \chi^{2'}(n, \lambda)$

**8** Using the moment generating function of the $\chi^{2'}(n, \lambda)$-distribution, derive its mean and variance.

**9** This exercise is another way to obtain the mean and variance of a $\chi^{2'}(n, \lambda)$-distribution.

**(a)** Let $x_i \sim N(\mu_i, 1), \lambda = \frac{1}{2}\sum_{i=1}^{n} \mu_i^2$. Show that $E\left(\sum_{i=1}^{n} x_i^2\right) = n + 2\lambda$.

**(b)** Observe that $x_i - \mu_i \sim \mathrm{N}(0, 1)$ and verify that

$$x_i^2 = (x_i - \mu_i)^2 + 2\mu_i(x_i - \mu_i) + \mu_i^2$$

**(c)** Use the moment generating function to show that the fourth central moment of a standard normal distribution is 3.

**(d)** Show that the variance

$$v(x_i^2) = 2 + 4\mu_i^2$$

(*Hint:* The first and third central moments of a standard normal distribution are zero.)

**(e)** Using the result in (d) show that

$$v\left(\sum_{i=1}^{n} x_i^2\right) = 2n + 8\lambda$$

**10** Derive the mean and variance of the $F'(n_1, n_2, \lambda)$ distribution.

**11** **(a)** Let $u_1 \sim \chi^{2'}(n_1, \lambda), u_2 \sim \chi^2_{n_2}$ and $u_1$ and $u_2$ are independent. Show that the joint distribution of $u_1$ and $u_2$ is

$$f(u_1, u_2) = \sum_{k=0}^{\infty} \alpha_k u_1^{\frac{1}{2}n_1+k-1} u_2^{\frac{1}{2}n_2-1} e^{-\frac{1}{2}(u_1+u_2)}$$

where

$$\alpha_k = \frac{e^{-\lambda}\lambda^k}{k!} \frac{1}{2^{\frac{1}{2}(n_1+n_2)+k}\Gamma(\frac{1}{2}n_1 + k)\Gamma(\frac{1}{2}n_2)}$$

**(b)** Let $u_1 = \frac{n_1 v z}{n_1 v + n_2}$ and $u_2 = \frac{n_2 z}{n_1 v + n_2}$. Find the joint distribution of $v$ and $z$.

**(c)** Find the marginal distribution of $v$ and show that the result is the non-central $F$-distribution with degrees of freedom $n_1$ and $n_2$ and non-centrality parameter $\lambda$.

**12** **(a)** Derive the mean and the variance of the $t_n$ and the central $F_{n_1,n_2}$-distribution.

**(b)** If the random variable $r$ is such that $r/(n\lambda + 1)$ has a central $F$-distribution with $a - 1$ and $a(n - 1)$ degrees of freedom show that

$$\widehat{\lambda} = \frac{1}{n}\left[r\left(1 - \frac{2}{a(n-1)}\right) - 1\right]$$

is an unbiased estimator of $\lambda$. (*Note*: In certain analysis of variance problems $r$ is a calculated $F$-statistic and $\lambda$ is a variance ratio.)

**13** Using Helmert's matrix of Section 1 show why $\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma^2$ has a $\chi^2_{n-1}$-distribution when $\mathbf{x}$ is $N(\mu\mathbf{1}, \sigma^2\mathrm{I})$.

**14** From the given definition of the $t_n$ and $\chi^2_n$-distributions show why $\frac{\bar{x}-\mu}{1/\sqrt{n}}\sqrt{\frac{n-1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t_{n-1}$.

**15** When $\mathbf{x}$ is $N(\boldsymbol{\mu}_1, \mathbf{I})$ and $\mathbf{y}$ is $N(\boldsymbol{\mu}_2, \mathbf{I})$ and the correlation matrix between $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{R}$, what are the mean and variance of $\mathbf{x}'\mathbf{A}\mathbf{y}$?

**16** A characterization of the multivariate normal distribution is that $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$ if and only if $\boldsymbol{\lambda}'\mathbf{x}$ has a univariate normal distribution. Using this as a definition of the multivariate normal distribution, derive its moment generating function from that of the univariate normal. (*Hint:* $\mathrm{M}_{\mathrm{X}}(\mathbf{t}) = \mathrm{M}_{\mathrm{t'x}}(1)$)

**17** If $u$ and $v$ have a bivariate normal distribution with zero means, show that

$$\mathrm{cov}(u^2, v^2) = 2[\mathrm{cov}(u, v)]^2.$$

**18** Suppose that $x \sim F(n_1, n_2)$ and $\Pr\{x \geq F_{n_1,n_2,\alpha}\} = \alpha$. Prove that $F_{n_2,n_1,1-\alpha} = 1/F_{n_1,n_2,\alpha}$.

**19** Show that if $t$ has a non-central $t$-distribution with $n$ degrees of freedom and non-centrality parameter $\lambda$ then $t^2$ has a non-central $F$-distribution with degrees of freedom 1 and $n$ and non-centrality parameter $\lambda$.

**20** Show that for a square symmetric $n$-dimensional matrix A, $\sum_{i=1}^{n}(\lambda_i \text{ of } \mathbf{A})^r = \mathrm{tr}A^r$.

**21** Let $y \sim N(0,\mathbf{I})$. Show that $\mathbf{y}'\mathbf{A}\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{A})\mathbf{y}$ are independent if and only if $\mathbf{A}$ is an idempotent matrix.

**22** Let $y \sim N(0,\mathbf{I}_n)$. Let $\mathbf{J}$ be and $n \times n$ matrix of ones. Show that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ and } \bar{x} \text{ are independent}$$

(*Hint:* Represent $\bar{x}$ and $s^2$ as quadratic forms using matrices $\mathbf{I}$ and $\mathbf{J}$.)

**23** Use equation (50) to show that equation (52) implies equation (53).

**24** **(a)** Verify (54) using the expression of bilinear forms as quadratic forms.
**(b)** Verify (57).

**25** Show that for non-singular $\mathbf{V}$, $\mathbf{AV}$ is idempotent if and only if $\mathbf{V}^{1/2}\mathbf{A}\,\mathbf{V}^{1/2}$ is idempotent. Thus, we have an alternative necessary and sufficient condition for Theorem 5 to hold true.

**26** This exercise is based on Exercise 17.2, p. 235 of Gruber (2014). Using the singular value decomposition find a transformation to express each of these quadratic forms as a sum of squares and write the quadratic form as a sum of squares.
Refer to Example 2 to see what is required.
**(a)** $5x_1^2 - 4x_1x_2 + 2x_2^2$
**(b)** $4x_1^2 + 2x_2^2 + 2x_3^2 + 4x_1x_2 + 4x_1x_3$

# 3

# REGRESSION FOR THE FULL-RANK MODEL

## 1. INTRODUCTION

### a. The Model

Regression analysis is designed for situations where a researcher thinks that a variable is related to one or more other measurements made, usually on the same object. A purpose of the analysis is to use data (observed values of the variables) to estimate the form of this relationship. An example would be to use information on income and the number of years of formal schooling (beyond the sixth grade) to estimate the extent to which a person's annual income is related to his/her years of schooling. One possibility is that for a person with zero years beyond sixth grade, a researcher would anticipate an income of $a$ dollars. For each year of schooling beyond sixth grade, a person has the researcher would expect that his/her income would be larger by $b$ dollars. Thus, for a person with $x$ years of schooling beyond sixth grade, the researcher would expect an annual income of $a + bx$ dollars. When we say that the researcher would expect an annual income of $a + bx$ dollars, we refer to the average income of all people that have had $x$ years of school beyond sixth grade. Some might make more and some might make less money but the researcher would expect the incomes to average out to $a + bx$ dollars. If $y$ denotes income and $x$ denotes years of schooling beyond sixth grade, we write $E(y)$ for expected income. This leads to the relationship

$$E(y) = a + bx. \tag{1}$$

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

The attempted description of how we think one variable is related to another variable is an example of what is called *model building*. The model here that a person's income is expected to be $a + bx$ where $x$ is his/her number of years of schooling beyond sixth grade is a *linear model* because we envisage $E(y)$ as being a linear combination of the unknowns $a$ and $b$. These unknowns are called *parameters*. There are of course endless other models, nonlinear in $a$ and $b$, that might be postulated. Some examples include $E(y)$ as a function of $a^x$ or $(\log x)^b$ or perhaps $b^x$. One way to guess what form of model might be reasonable would be to plot data points of income versus years of schooling beyond sixth grade on a graph where income is the vertical ($y$-axis) and years of schooling beyond sixth grade is the horizontal ($x$-axis). For the data points in Example 1 to be described in Section 1c, see the scatterplot in Figure 3.1.

The points appear to lie reasonably close to a straight line. The line in the diagram is the "best fitting" line for these data points. In Section 1c, we shall explain how to find this line and why it is the best fitting.

The linear model is the one that has received greatest attention in theory and practice. From the theoretical point of view, it is mathematically tractable. It has shown itself to be of great value in a wide variety of practical applications to diverse areas of knowledge. Many nonlinear models can be rearranged in a linear form,

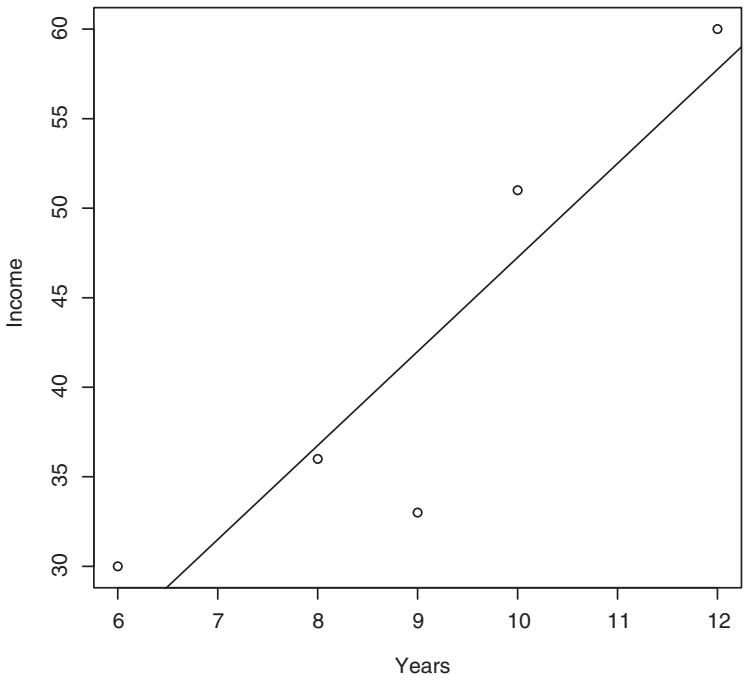



**FIGURE 3.1** A Scatterplot of Income vs. Years of School Beyond Sixth Grade with Best Fitting Line

usually by making a transformation on either the $x$ or $y$ variable or both. Modern computing methods make analyses of data using linear models ever more attainable.

Equation (1) is the equation of our model, in this case the model of how expected income and years of schooling are related. The equation is not the whole model. Its other parts have yet to be described. Since the model is something being conjectured, $a$ and $b$ can never be known. The best that can be done is to obtain estimates from data that we assume to be a random sample from some population to which we believe our equation applies. The model is called a regression model. Since its equation is linear, it is more correctly called *linear regression.* The variable denoted by $y$ is called the *dependent variable.* Correspondingly, $x$ is called the *independent variable.*

### b. Observations

In gathering data, the income of every person with $x$ years of schooling beyond sixth grade will not be exactly $a + bx$ (with $a$ and $b$ being the same for each person). Indeed, this fact was recognized by writing the equation of the model as $E(y) = a + bx$ rather than $y = a + bx$. Thus, if $y_i$ is the income of a person with $x_i$ years of schooling, we write

$$E(y_i) = a + bx_i, \tag{2}$$

where $E(y_i)$ is not the same as $y_i$. The difference $y_i - E(y_i)$ represents the deviation of the observed value $y_i$ from its expected value $E(y_i)$ and is written as

$$e_i = y_i - E(y_i) = y_i - a - bx_i. \tag{3}$$

Hence

$$y_i = a + bx_i + e_i, \tag{4}$$

which we now take as the equation of the model.

The deviation $e_i$ defined in (3) represents the extent to which an observed $y_i$ differs from its expected value given in equation (2). Moreover equations (2), (3), and (4) apply to each of our $N$ observations $y_1, y_2, \ldots, y_N$. Thus the $e_i$'s include all manner of discrepancies between observed $y$'s and their expected values. For example, they include measurement errors in $y_i$. Its recorded value might not be exactly what the person's income is. They include deficiencies in the model itself, the extent to which $a + bx_i$ is in fact not the person's income. Variables other than years of schooling might affect it, the person's age, for example. As a result, the $e$'s are considered random variables. They are usually called random errors or random residuals.

In order to complete the description of the model in terms of equation (4) the characteristics of the $e$'s must be specified. Customary specifications are

1. The expected value of $e_i$ are zero;
2. The variance of $e_i$ is $\sigma^2$ for all $i$;
3. The covariance between any pairs of the $e$'s is zero.

Expressing this information in terms of equations we have that

$$E(e_i) = 0, \tag{5}$$

as is obvious from the definition in (3),

$$v(e_i) = E[e_i - E(e_i)]^2 = \sigma^2, \quad \text{for all } i, \tag{6}$$

and

$$\text{cov}(e_i e_j) = E[(e_i - E(e_i))(e_j - E(e_j))] = 0, \text{ for all } i \neq j. \tag{7}$$

Equations (2)–(7) now constitute the model. They form the basis for the procedure used for estimating $a$ and $b$.

### c. Estimation

There are several well-recognized methods that can be used for estimating $a$ and $b$. A frequently used method is known as least squares.

Least-squares estimation involves minimizing the sum of the squares of the deviations of the observed $y_i$'s from their expected values. In Figure 1 the points on the line right above the observed values are the expected values for $y_i$. In view of (3) this sum of squares is,

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - E(y_i)]^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2. \tag{8}$$

Although $a$ and $b$ are fixed (but unknown values), for the moment, think of them as mathematical variables. Those values of $a$ and $b$ that minimize (8) are the least-square estimators of $a$ and $b$. They will be denoted by $\hat{a}$ and $\hat{b}$. To minimize (8) we differentiate it with respect to $a$ and $b$ and equate the derivatives to zero. Thus from (8)

$$\frac{\partial(\mathbf{e}'\mathbf{e})}{\partial a} = -2\sum_{i=1}^{N}(y_i - a - bx_i) = -2\left(\sum_{i=1}^{N} y_i - Na - b\sum_{i=1}^{N} x_i\right) = 0 \tag{9}$$

and

$$\frac{\partial(\mathbf{e}'\mathbf{e})}{\partial b} = -2\sum_{i=1}^{N} x_i(y_i - a - bx_i) = -2\left(\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i - b\sum_{i=1}^{N} x_i^2\right) = 0. \tag{10}$$

Rewriting (9) and (10), we obtain the normal equations

$$N\hat{a} + \hat{b}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \quad \text{and} \quad \hat{a}\sum_{i=1}^{N} x_i + \widehat{b}\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \tag{11}$$

Using the dot notation

$$x. = \sum_{i=1}^{N} x_i \quad \text{and} \quad y. = \sum_{i=1}^{N} y_i, \tag{12}$$

and the corresponding bar notation for observed means,

$$\bar{x} = \frac{x.}{N} \quad \text{and} \quad \bar{y} = \frac{y.}{N}, \tag{13}$$

the least-square estimators obtained as the solution to (11) may be written

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{N \sum_{i=1}^{N} x_i y_i - x.y.}{N \sum_{i=1}^{N} x_i^2 - x_.^2} \tag{14}$$

and

$$\hat{a} = \bar{y}. - \hat{b}\bar{x} = \frac{(y. - \hat{b}x.)}{N}. \tag{15}$$

Although the two expressions in (14) and (15) are algebraically equivalent, it is better to use the last expression in (14) and (15) because division does not happen until the end of the calculation. This lessens the possibility of a round off error. In many regression computations, $\hat{b}$ is included in the formula so it is important to include at least four decimal places in the answer. Otherwise, false inferences can result based on the least-square estimators.

**Example 1 An Example of a Least-square Fit** Suppose that in a sample of five persons, their incomes (in thousands of dollars) and years of schooling are as follows:

| $i$ Person | $y_i$ (Income, \$1000) | $x_i$ (Years of Schooling Beyond Sixth Grade) |
|---|---|---|
| 1 | 30 | 6 |
| 2 | 60 | 12 |
| 3 | 51 | 10 |
| 4 | 36 | 8 |
| 5 | 33 | 9 |
| $N = 5$ | $y. = 210$ | $x. = 45$ |
| | $\bar{y}. = 42$ | $\bar{x}. = 9$ |
| | $\sum_{i=1}^{5} y_i^2 = 9486$ | $\sum_{i=1}^{5} x_i^2 = 425$ |
| | | $\sum_{i=1}^{5} x_i y_i = 1995.$ |

From (11), the normal equations for obtaining $\hat{a}$ and $\hat{b}$ are

$$5\hat{a} + 45\hat{b} = 210 \quad \text{and} \quad 45\hat{a} + 425\hat{b} = 1995.$$

From (14) and (15) the solutions are

$$\hat{b} = \frac{5(1995) - (45)(210)}{5(425) - 45^2} = 5.25$$

and

$$\hat{a} = \frac{210 - 5.25(45)}{5} = -5.25.$$

Hence the estimated regression equation is

$$\widehat{E(y_i)} = \hat{a} + \hat{b}x_i = -5.25 + 5.25x_i,$$

where the "hat" over $E(y_i)$ denotes "estimator of" $E(y_i)$ just as does $\hat{a}$ of $a$.

A simple program to do the regression using R would be

```
> years=c(6,12,10,8,9)
> income=c(30,60,51,36,33)
> lm(income~years)
```

The output would be

```
Call:
lm(formula = income ~ years)

Coefficients:
(Intercept)        years
      -5.25         5.25
```

□

### d. The General Case of $k$ $x$ Variables

Suppose that in the study of annual income and years of schooling, we also consider the person's age as a factor affecting income. The model envisaged in (1) is now extended to be

$$E(y_i) = a + b_1x_1 + b_2x_2,$$

where $x_1$ represents years of schooling and $x_2$ is age. Thus, for the $i$th person in our data who has had $x_{i1}$ years of schooling and whose age is $x_{i2}$, equation (4) could be replacing $a$ with $b_0$

$$y_i = b_0 + b_1x_{i1} + b_2x_{i1} + e_i, \tag{16}$$

for $i = 1, 2,\ldots, N$.

Now define the following matrix and vectors:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Then the complete set of equations represented by (16) is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \text{ with } E(\mathbf{y}) = \mathbf{Xb}. \tag{17}$$

Extension to more than just 2 $x$ variables is clear. For $k$ variables

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \tag{18}$$

and $\mathbf{y}$ and $\mathbf{e}$ defined as above are unchanged. Equation (17) is also unchanged. It represents the model no matter how many $x$ variables there are so long as the number $k$ of $x$ variables are less than the number of observations $N$. More formally, $k < N$. The model in (17) will be studied throughout this book. There will be many different variations of it depending on the $\mathbf{X}$ matrix. When $k \geq N$, the values of the $b_i$ can be derived so that $\mathbf{y} = \mathbf{Xb}$ exactly, and there is no estimation problem.

Complete specification of the model demands that we define the distributional properties of the vector $\mathbf{e}$. For the moment, we only need its expected value and its variance. In accord with (5), (6), and (7) we have that

$$E(\mathbf{e}) = \mathbf{0} \text{ and } \text{var}(\mathbf{e}) = E[\mathbf{e} - E(\mathbf{e})][\mathbf{e} - E(\mathbf{e})]' = E(\mathbf{ee}') = \sigma^2\mathbf{I}_N. \tag{19}$$

The derivation of the least-square estimators does not require us to specify the mathematical form of the distribution of $\mathbf{e}$, for example, whether it is normal, exponential, or some other distribution. However, we will need to specify this sort of information later in order to consider hypothesis testing and confidence intervals.

Derivation of the least-square estimator of $\mathbf{b}$ follows the same procedure used in establishing (11), namely minimization of the sum of squares of observations from their expected values. Similar to (8), this sum of squares with $E(\mathbf{e}) = 0$ and hence $E(\mathbf{y}) = \mathbf{Xb}$, is

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= [\mathbf{y} - E(\mathbf{y})]'[\mathbf{y} - E(\mathbf{y})] = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}. \end{aligned}$$

In order to obtain the estimator $\hat{\mathbf{b}}$, that value of $\mathbf{b}$ that minimizes $\mathbf{e}'\mathbf{e}$, we must differentiate $\mathbf{e}'\mathbf{e}$ with respect to the elements of $\mathbf{b}$ and setting the result equal to zero [See, for example, Section 8.5 of Searle (1966) or Section 4.7 of Gruber (2014)]. We get

$$\frac{\partial(\mathbf{e}'\mathbf{e})}{\partial\mathbf{b}} = -2\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}.$$
$$-2\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}.$$

and then

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{20}$$

The equations (20) are known as the *normal equations.* Provided that $(\mathbf{X}'\mathbf{X})^{-1}$ exists, they have a unique solution for $\hat{\mathbf{b}}$,

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{21}$$

Here is where the description "full-rank model" applies. When $\mathbf{X}'\mathbf{X}$ is non-singular (of full rank) the unique solution of (20) can be written as (21). On the other hand, when $\mathbf{X}'\mathbf{X}$ is singular, the solution will take the form

$$\hat{\mathbf{b}} = \mathbf{G}\mathbf{X}'\mathbf{y}, \tag{22}$$

where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. This solution is not unique because generalized inverses are not unique as was pointed out in Chapter 1. Finding least-square estimators for the non-full-rank case will be taken up in Chapter 5.

By the nature of $\mathbf{X}$, as shown in (18), $\mathbf{X}'\mathbf{X}$ is square of order $k+1$ with elements that are sums of squares and products and $\mathbf{X}'\mathbf{y}$ is the vector of sums of products of the observed $x$'s and $y$'s. As a result, we have,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{N} x_{ik}y_i \end{bmatrix}. \tag{23}$$

and

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & x.1 & x.2 & \cdots & x.k \\ x.1 & \sum_{i=1}^{N} x_{i1}^2 & \sum_{i=1}^{N} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{N} x_{i1}x_{ik} \\ x.2 & \sum_{i=1}^{N} x_{i1}x_{i2} & \sum_{i=1}^{N} x_{i2}^2 & \cdots & \sum_{i=1}^{N} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ x.k & \sum_{i=1}^{N} x_{i1}x_{ik} & \sum_{i=1}^{N} x_{i2}x_{ik} & \cdots & \sum_{i=1}^{N} x_{ik}^2 \end{bmatrix}. \tag{24}$$

**Example 2** Suppose that in Example 1 the ages of the persons supplying the data had been available.

| $i$ (Person) | $y_i$ (Income, \$1000) | $x_{i1}$ (Years of Schooling Beyond Sixth Grade) | $x_{i2}$ (Age) |
|---|---|---|---|
| 1 | 30 | 6 | 28 |
| 2 | 60 | 12 | 40 |
| 3 | 51 | 10 | 32 |
| 4 | 36 | 8 | 36 |
| 5 | 33 | 9 | 34 |
| $N = 5$ | $y_{\cdot} = 210$ | $x_{\cdot 1} = 45$ | $x_{\cdot 2} = 170$ |
| | $\bar{y}_{\cdot} = 42$ | $\bar{x}_{\cdot 1} = 9$ | $\bar{x}_{\cdot 2} = 34$ |

$$\sum_{i=1}^{5} y_i^2 = 9486 \quad \sum_{i=1}^{5} x_{i1}^2 = 425 \quad \sum_{i=1}^{5} x_{i2}^2 = 5860$$

$$\sum_{i=1}^{5} x_{i1}x_{i2} = 1562 \quad \sum_{i=1}^{5} x_{i1}y_i = 1995 \quad \sum_{i=1}^{5} x_{i2}y_i = 7290$$

Putting these values into (24), we have

$$\mathbf{X} = \begin{bmatrix} 1 & 6 & 28 \\ 1 & 12 & 40 \\ 1 & 10 & 32 \\ 1 & 8 & 36 \\ 1 & 9 & 34 \end{bmatrix}, \quad \mathbf{X}'\mathrm{X} = \begin{bmatrix} 5 & 45 & 170 \\ 45 & 425 & 1562 \\ 170 & 1562 & 5860 \end{bmatrix} \tag{25}$$

and

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{2880}\begin{bmatrix} 50656 & 1840 & -1960 \\ 1840 & 400 & -160 \\ -1960 & -160 & 100 \end{bmatrix}. \tag{26}$$

From (23),

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 210 \\ 1995 \\ 7290 \end{bmatrix}. \tag{27}$$

Equation (21) gives

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \frac{1}{2880}\begin{bmatrix} 50656 & 1840 & -1960 \\ 1840 & 400 & -160 \\ -1960 & -160 & 100 \end{bmatrix}\begin{bmatrix} 210 \\ 1995 \\ 7290 \end{bmatrix} \\ &= \begin{bmatrix} 7 \\ 6.25 \\ -0.625 \end{bmatrix}. \end{aligned} \tag{28}$$

A simple program in R to do this problem would be

```
> income=c(30,60,51,36,33)
> school=c(6,12,10,8,9)
> age=c(28,40,32,36,34))
> lm(income~school+age)
```

The resulting output is

```
lm(formula = income ~ years + age)

Coefficients:
(Intercept)        years          age
      7.000        6.250       -0.625
```

Thus, from these data, the estimated form of the relationship between $y$ and $x_1$ and $x_2$ is

$$\widehat{E(y)} = 7.000 + 6.250x_1 - 0.625x_2.$$

□

### e. Intercept and No-Intercept Models

When all of the $x$'s are zero in the above models, $E(y_0) = b_0$ with estimator $\hat{b}_0$. Thus, for $x_1 = x_2 = 0$ in Example 2, the estimated value of $E(y)$ is $b_0 = 7.000$. Models of

this kind are called *intercept models*. The intercept is $b_0$, the value of $E(y)$ when all $x$'s are zero.

Sometimes, it is appropriate to have no term $b_0$ in the model. In this case, the model is called a *no-intercept model.* The matrix $\mathbf{X}$ has no vector of 1's in it as does $\mathbf{X}$ of (25), for example, and $\mathbf{X}'\mathbf{X}$ is then the matrix of sums of squares and products of the observations without the first row and column of totals as seen in (24).

**Example 3 A No-Intercept Model** For the data of Example 2 for the no-intercept model,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 425 & 1562 \\ 1562 & 5860 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 1995 \\ 7290 \end{bmatrix}.$$

The least-square estimators are given by

$$\begin{aligned} \hat{\mathbf{b}} &= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \frac{1}{50656}\begin{bmatrix} 5860 & -1562 \\ -1562 & 425 \end{bmatrix}\begin{bmatrix} 1995 \\ 7290 \end{bmatrix} = \begin{bmatrix} 5.9957 \\ -0.3542 \end{bmatrix}. \end{aligned} \tag{29}$$

Thus, the no-intercept model leads to $\widehat{E(y)} = 5.9957x_1 - 0.3542x_2$.

Redoing this computation using R we have

```
> Income=c(30,60,51,36,33)
> School=c(6,12,10,8,9)
> Age=c(28,40,32,36,34)
> lm(Income~-1+School+Age)
```

The resulting output is

```
Call:
lm(formula = income ~ years + age - 1)

Coefficients:
  years      age
 5.9957  -0.3542
```

□

## 2. DEVIATIONS FROM MEANS

The matrix $\mathbf{X}'\mathbf{X}$ and matrix $\mathbf{X}'\mathbf{y}$ as shown in (23) and (24) have as elements the sums of squares and products of the observations. However, it is well known that the regression coefficients $b_1, b_2, \ldots, b_k$ can be estimated using a matrix and vector that are just like $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$, only involving sums of squares and products corrected for their means. We establish this formulation. Some additional notation is needed.

If we define

$$\mathbf{1}_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_1 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix}, \tag{30}$$

$\mathbf{X}$ can be written

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_1 \end{bmatrix}, \tag{31}$$

where the order of $\mathbf{1}$ is $N \times 1$ and as in (30), $\mathbf{X}_1$ is the $N \times k$ matrix of the observed $x$'s. In addition, define

$$\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_{.1} & \bar{x}_{.2} & \ldots & \bar{x}_{.k} \end{bmatrix} \tag{32}$$

as the vector of means of the observed $x$'s. These definitions imply

$$\mathbf{1}_N'\mathbf{1}_N = N, \mathbf{1}'\mathbf{y} = N\bar{y} \quad \text{and} \quad \mathbf{1}'\mathbf{X}_1 = N\bar{\mathbf{x}}', \tag{33}$$

where for convenience we write $\bar{y}$ in place of $\bar{y}.$ for the mean.

Using (33) we may express the solution $\hat{\mathbf{b}}$ as

$$\begin{aligned} \hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[ \begin{bmatrix} \mathbf{1}' \\ \mathbf{X}_1' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{X}_1 \end{bmatrix} \right]^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{X}_1' \end{bmatrix} \mathbf{y} \\ &= \begin{bmatrix} N & N\bar{\mathbf{x}}' \\ N\bar{\mathbf{x}} & \mathbf{X}_1'\mathbf{X}_1 \end{bmatrix}^{-1} \begin{bmatrix} N\bar{y} \\ \mathbf{X}_1'\mathbf{y} \end{bmatrix}. \end{aligned}$$

Using the procedure for inverting a partitioned symmetric matrix given in equation (55) of Section 1.7 or in Section 3.5 of Gruber (2014), this becomes

$$\hat{\mathbf{b}} = \begin{bmatrix} \frac{1}{N} + \bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\bar{\mathbf{x}} & \mathbf{S}^{-1} \end{bmatrix} \begin{bmatrix} N\bar{y} \\ \mathbf{X}_1'\mathbf{y} \end{bmatrix}, \tag{34}$$

where

$$\mathbf{S} = \mathbf{X}_1'\mathbf{X}_1 - N\overline{\mathbf{x}\mathbf{x}}'. \tag{35}$$

Then on partitioning

$$\tilde{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} b_0 \\ \boldsymbol{\mathscr{b}} \end{bmatrix},$$

(34) can be written as

$$\begin{bmatrix} b_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \left[ \begin{bmatrix} \frac{1}{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\bar{\mathbf{x}}' \\ \mathbf{I} \end{bmatrix} \mathbf{S}^{-1} \begin{bmatrix} -\bar{\mathbf{x}} & \mathbf{I} \end{bmatrix} \right] \begin{bmatrix} N\bar{y} \\ \mathbf{X}_1'\mathbf{y} \end{bmatrix}$$
$$= \begin{bmatrix} \bar{y} - \bar{\mathbf{x}}'\mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{y} - N\bar{y}\bar{\mathbf{x}}) \\ \mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{y} - N\bar{y}\bar{\mathbf{x}}) \end{bmatrix}$$

so that

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{y} - N\bar{y}\bar{\mathbf{x}}) \tag{36}$$

and

$$\hat{b}_0 = \bar{y} - \bar{\mathbf{x}}'\hat{\mathbf{b}}. \tag{37}$$

Now consider $\mathbf{S}$ given in (35). First,

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{bmatrix} \sum_{i=1}^{k} x_{i1}^2 & \sum_{i=1}^{k} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{k} x_{i1}x_{ik} \\ \sum_{i=1}^{k} x_{i1}x_{i2} & \sum_{i=1}^{k} x_{i2}^2 & \cdots & \sum_{i=1}^{k} x_{i2}x_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{k} x_{i1}x_{ik} & \sum_{i=1}^{k} x_{i2}x_{ik} & & \sum_{i=1}^{k} x_{ik}^2 \end{bmatrix}, \tag{38}$$

the matrix of sums of squares and products of the $k$ observed $x$ variables. Second, by the nature of $\bar{\mathbf{x}}$ in (32), the matrix $N\overline{\mathbf{x}\mathbf{x}}'$ is

$$N\overline{\mathbf{x}\mathbf{x}}' = \{N\bar{x}_{.p}\,\bar{x}_{.q}\} \text{ for } p, q = 1, 2, \ldots, k.$$

Thus

$$\mathbf{S} = \left\{ \sum_{i=1}^{k} x_{ip}x_{iq} - N\bar{x}_{.p}\bar{x}_{.q} \right\}, \text{ for } p, q = 1, 2, \ldots, k.$$

Define

$$\mathcal{X} = \mathbf{X}_1 - \mathbf{1}_N\bar{\mathbf{x}}' \tag{39}$$

as the matrix of observed $x$'s expressed as deviations from their means. Then it is easily shown that $\mathbf{S}$ as just derived is

$$\mathbf{S} = \mathcal{X}'\mathcal{X}. \tag{40}$$

Thus, the matrix $\mathbf{S}$ in (36) is that of the corrected sums of squares and products of the $x$'s. In a like manner, the other term in (36) is

$$\begin{aligned}\mathbf{X}_1'\mathbf{y} - N\bar{y}\bar{\mathbf{x}} &= \left\{ \sum_{i=1}^{k} x_{ip}y_i - N\bar{x}_{.p}\bar{y} \right\}^{b}, \text{ for } p = 1, 2, \ldots, k, \\ &= \tilde{\mathcal{X}}'\mathbf{y}.\end{aligned}$$

This is the vector of corrected sums and products of the $x$'s and $y$'s. Hence, just as $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (21), we can now write from (36),

$$\hat{\boldsymbol{\theta}} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y}. \tag{41}$$

The $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ is the inverted matrix of corrected sums of squares and products of the $x$'s pre-multiplying the vector of corrected sums of products of the $x$'s and the $y$'s. Then as in (37), $\hat{b}_0$ is given by

$$\hat{b}_0 = \bar{y} - \hat{\boldsymbol{\theta}}'\bar{\mathbf{x}}. \tag{42}$$

The results (41) and (42) are the familiar expressions for calculating the regression estimators using corrected sums of squares.

**Example 4 Regression Calculations Using the Corrected Sums of Squares and Products** From the data of Example 2, we have

$$\mathcal{X}' = \begin{bmatrix} -3 & -6 \\ 3 & 6 \\ 1 & -2 \\ -1 & 2 \\ 0 & 0 \end{bmatrix}, \tag{43}$$

$$\mathcal{X}'\mathcal{X} = \begin{bmatrix} 20 & 32 \\ 32 & 80 \end{bmatrix},$$

$$(\mathcal{X}'\mathcal{X})^{-1} = \frac{1}{144}\begin{bmatrix} 20 & -8 \\ -8 & 5 \end{bmatrix}, \tag{44}$$

$$\text{and } \mathcal{X}'\mathbf{y} = \begin{bmatrix} 105 \\ 150 \end{bmatrix}. \tag{45}$$

Then

$$\hat{\boldsymbol{\theta}} = \frac{1}{144}\begin{bmatrix} 20 & -8 \\ -8 & 5 \end{bmatrix}\begin{bmatrix} 105 \\ 150 \end{bmatrix} = \begin{bmatrix} 6.250 \\ -0.625 \end{bmatrix}$$

as in (28). From (42),

$$\hat{b}_0 = 42 - [6.250 \quad -0.625] \begin{bmatrix} 9 \\ 34 \end{bmatrix} = 7.000$$

as in (28). Derivation of $\tilde{\mathbf{b}}$ in this manner does not apply for the no-intercept model which contains no $b_0$ term. For then the partitioning of $\mathbf{b}'$ as $\begin{bmatrix} b_0 & \tilde{\mathbf{b}}' \end{bmatrix}$ does not exist. This is because $\mathbf{b}'$ is itself the vector of the $b$'s corresponding to the $k$ $x$ variables and $\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ is based on the uncorrected sum of squares and products as exemplified in (24).

## 3. SOME METHODS OF ESTIMATION

In this section, we summarize four different ways to obtain the least-square estimator and a method of obtaining an alternative to the least-square estimator that is useful for certain types of data that arise in applications. In obtaining the least-square estimator, we shall assume a model of the form $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ where $\mathbf{X}$ has full column rank, $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$, and $E(\boldsymbol{e}) = \mathbf{0}$. The non-full-rank case will be considered in Chapter 5. To obtain an alternative to the least-square estimator, we shall assume that $\mathbf{b}$ is a random variable with a known mean and covariance matrix.

### a. Ordinary Least Squares

This involves choosing $\hat{\mathbf{b}}$ as the value of $\mathbf{b}$ which minimizes the sum of squares of observations from their expected values. More formally, chose $\tilde{\mathbf{b}}$ as that $\mathbf{b}$ that minimizes

$$\sum_{i=1}^{N} [y_i - E(y)]^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

The resulting estimator is as we have seen

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

### b. Generalized Least Squares

This is also called weighted least squares. Assume that the variance covariance matrix of $\mathbf{e}$ is $\text{var}(\mathbf{e}) = \mathbf{V}$. Now minimize $(\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})$ with respect to $\mathbf{b}$. The resulting estimator is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

When it is assumed that the components of var(**e**) are equal and uncorrelated, that is, $\mathbf{V} = \sigma^2\mathrm{I}$, the generalized or weighted least-square estimator reduces to the ordinary least-square estimator.

### c. Maximum Likelihood

The derivation of the least-square estimator made no assumption about the form of the distribution of the error term. The likelihood is the joint distribution of the errors. We shall assume that this distribution is multivariate normal with zero mean and covariance matrix $\mathbf{V}$. Then $\mathbf{y} - \mathbf{Xb} \sim N(\mathbf{0}, \mathbf{V})$. The likelihood function is

$$L = (2\pi)^{-\frac{1}{2}N} |\mathbf{V}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}\right\}.$$

The maximum likelihood estimator is the $\mathbf{b}$ that maximizes the likelihood. We can maximize the $\log(L)$ where

$$\log L = -\frac{1}{2}N \log(2\pi) - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{Xb}).$$

by minimizing $(\mathbf{Y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{Xb})$, thus obtaining the generalized least-square estimator,

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

If we assume that $\mathbf{V} = \sigma^2\mathbf{I}$, this reduces to the ordinary least-square estimator.

Two well-known points are worth emphasizing about these estimators

(i) Least-square estimation does not pre-suppose any distributional properties about the $e$'s other than zero means and finite variances.

(ii) Maximum likelihood estimation under normality assumptions leads to the generalized least-square estimators and, when $\mathbf{V} = \sigma^2\mathbf{I}$, the ordinary least-square estimators.

### d. The Best Linear Unbiased Estimator (b.l.u.e.)(Gauss–Markov Theorem)

We shall show that the least-square estimator is the linear unbiased estimator of the parameters of a regression that has minimum variance (the best linear unbiased estimator b.l.u.e). This is the content of the Gauss–Markov theorem.

For any row vector $\mathbf{t}'$ with the same number of columns as there are rows of $\mathbf{b}$, the scalar $\mathbf{t}'\mathbf{b}$ is a linear function of the elements of the vector of parameters $\mathbf{b}$.

Three characteristics of the estimator under study are linearity, unbiasedness, and being the best estimator (the one with the smallest variance). We shall clarify these characteristics.

(i) *Linearity:* The estimator is to be a linear function of the observations $\mathbf{y}$. Let this estimator be $\boldsymbol{\lambda}'\mathbf{y}$ where $\boldsymbol{\lambda}'$ is a row vector of order N. We shall show that $\boldsymbol{\lambda}$ is uniquely determined by the other two characteristics of the estimator.

(ii) *Unbiasedness:* The estimator $\boldsymbol{\lambda}'\mathbf{y}$ is to be unbiased for $\mathbf{t}'\mathbf{b}$. Therefore, we must have that $E(\boldsymbol{\lambda}'\mathbf{y}) = \mathbf{t}'\mathbf{b}$. However, $E(\boldsymbol{\lambda}'\mathbf{y}) = \boldsymbol{\lambda}'\mathbf{X}\mathbf{b}$ so that $\boldsymbol{\lambda}'\mathbf{X}\mathbf{b} = \mathbf{t}'\mathbf{b}$. Since this must be true for all $\mathbf{b}$, we have that

$$\boldsymbol{\lambda}'\mathbf{X} = \mathbf{t}'. \tag{46}$$

(iii) *A "best" estimator:* Here, "best" means that in the class of linear, unbiased estimators of $\mathbf{t}'\mathbf{b}$, the best is the one that has minimum variance. This is the criterion for deriving $\boldsymbol{\lambda}'$.

We now state and prove the Gauss–Markov theorem.

**Theorem 1** Assume that for the linear model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

$\text{var}(\mathbf{e}) = \mathbf{V}$. Then the best linear unbiased estimator of $\mathbf{t}'\mathbf{b}$ is

$$\mathbf{t}'\hat{\mathbf{b}} = \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

*Proof.* Since var $(\mathbf{e}) = \mathbf{V}$, var $(\mathbf{y}) = \mathbf{V}$. Then $\text{var}(\boldsymbol{\lambda}'\mathbf{y}) = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda}$. We must minimize this quantity with respect to the constraint ($\boldsymbol{\lambda}'\mathbf{X} = \mathbf{t}'$) in (46). To do this, we use the method of Lagrange multipliers (See Sections 22 and 23 of Gruber (2014) or any book on multivariable calculus.). Using $2\boldsymbol{\theta}$ as a vector of Lagrange multipliers, we therefore minimize

$$w = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} - 2\boldsymbol{\theta}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{t})$$

with respect to the elements of $\boldsymbol{\lambda}'$ and $\boldsymbol{\theta}'$. We differentiate $w$ with respect to $\boldsymbol{\theta}$, set it equal to zero and get (46). Differentiation of $w$ with respect to $\lambda$ gives

$$\mathbf{V}\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\theta} \quad \text{or} \quad \boldsymbol{\lambda} = \mathbf{V}^{-1}\boldsymbol{\theta},$$

since $\mathbf{V}^{-1}$ exists. Substitution in (46) gives $\mathbf{t}' = \boldsymbol{\lambda}'\mathbf{X} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ and so $\boldsymbol{\theta}' = \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$.

Hence,

$$\boldsymbol{\lambda}' = \boldsymbol{\theta}'\mathbf{X}'\mathbf{V}^{-1} = \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}. \tag{47}$$

From (45) we have that the b.l.u.e. of $\mathbf{t}'\mathbf{b}$ is

$$\mathbf{t}'\hat{\mathbf{b}} = \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

■

We have shown that the b.l.u.e. is the weighted or generalized least-square estimator. Its variance is

$$\text{var}(\mathbf{t}'\hat{\mathbf{b}}) = \mathbf{t}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{t}. \tag{48}$$

Since (47) is the sole solution to the problem of minimizing $\text{var}(\boldsymbol{\lambda}'\mathbf{y}) = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda}$ subject to the constraint (46), the b.l.u.e. $\boldsymbol{\lambda}'\mathbf{y}$ of $\mathbf{t}'\mathbf{b}$ is the unique estimator of $\mathbf{t}'\mathbf{b}$ having the properties of linearity, unbiasedness, and "bestness"—minimum variance of all linear unbiased estimators. Thus, the b.l.u.e. of $\mathbf{t}'\mathbf{b}$ is unique $\boldsymbol{\lambda}'\mathbf{y}$ for $\boldsymbol{\lambda}'$ as given in (47). Furthermore, this result is true for any vector $\mathbf{t}'$. Thus, for some other vector, say $\mathbf{p}'$, the b.l.u.e of $\mathbf{p}'\mathbf{b}$ is $\mathbf{p}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and its variance is $\mathbf{p}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{p}$. it may readily be shown that that the covariance of $\mathbf{p}'\hat{\mathbf{b}}$ with $\mathbf{t}'\hat{\mathbf{b}}$ is $\mathbf{p}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{t}$ (see Exercise 8).

Suppose that $\mathbf{t}'$ takes the value $\mathbf{u}_i'$, the $i$'th row of $\mathrm{I}_\mathrm{K}$. Then $\mathbf{u}_i'\mathbf{b} = b_i$, the $i$'th element of $\mathbf{b}$. The b.l.u.e. of $b_i$ is $\mathbf{u}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, the $i$'th element of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

Its variance is $\mathbf{u}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{u}_i$, the $i$'th diagonal term of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. Thus, by letting $\mathbf{t}'$ be in turn each row of $\mathbf{I}_\mathrm{K}$ the b.l.u.e. of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{49a}$$

with variance

$$\text{var}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{49b}$$

This expression for $\hat{\mathbf{b}}$ is the same as that given earlier. The generalized least-square estimator, the maximum likelihood estimator under normality assumptions and the b.l.u.e. are all the same $\hat{\mathbf{b}}$.

It has been shown above that $(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the b.l.u.e. of $\mathbf{b}$ when $\mathbf{V} = \mathbf{I}\sigma^2$. More generally, Mc Elroy (1967) has shown that $\hat{\mathbf{b}}$ is the b.l.u.e. of $\mathbf{b}$ whenever $\mathbf{V} = [(1-\boldsymbol{\rho})\mathbf{I} + \mathbf{1}\mathbf{1}'\boldsymbol{\rho}]\sigma^2$ for $0 \leq \rho < 1$. This form of $\mathbf{V}$ demands equality of the variances of the $e_i$'s and equality of the covariances between them, with the correlation between any two $e_i$'s being $\rho$. The case $\mathbf{V} = \mathbf{I}\sigma^2$ is obtained by setting $\rho = 0$.

### e. Least-squares Theory When The Parameters are Random Variables

In this section, we assume that the parameters of the regression models are random variables with a known mean and variance. We then show how to find the best linear estimator of a random variable $\mathbf{p}'\mathbf{b}$. The methodology used is that of Rao (1973, p. 274) using notation consistent with what we have used so far in this book.

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{50a}$$

together with the assumptions

$$E(\mathbf{b}) = \theta \text{ and } \operatorname{var}(\mathbf{b}) = \mathbf{F} \tag{50b}$$

where $\theta$ is a $k$-dimensional vector and $\mathbf{F}$ is a $k \times k$ positive definite matrix.

Also assume that

$$E(\mathbf{e}|\mathbf{b}) = \mathbf{0} \text{ and } \operatorname{var}(\mathbf{e}|\mathbf{b}) = \mathbf{V}. \tag{50c}$$

The following formulae connect the conditional and unconditional means and variances.

$$E(\mathbf{Y}) = E(E(\mathbf{Y}|\mathbf{e}), \tag{51a}$$

$$\begin{aligned} \operatorname{var}(\mathbf{Y}) &= E[\operatorname{var}(\mathbf{Y}|\mathbf{b})] + \operatorname{var}[E(\mathbf{Y}|\mathbf{b}) \\ &= \mathbf{V} + \mathbf{X}\mathbf{F}\mathbf{X}' \end{aligned} \tag{51b}$$

and

$$\begin{aligned} \operatorname{cov}(\mathbf{Y}, \mathbf{p}'\mathbf{b}) &= E[C(\mathbf{Y}, \mathbf{p}'\mathbf{b})|\mathbf{b}] + C[E(\mathbf{Y}|\mathbf{b})|\mathbf{p}'\mathbf{b}] \\ &= \mathbf{X}\mathbf{F}\mathbf{p}. \end{aligned} \tag{51c}$$

The objective is to determine a linear function $a + \mathbf{L}'\mathbf{Y}$ such that

$$E(\mathbf{p}'\mathbf{b} - a - \mathbf{L}'\mathbf{Y}) = 0 \tag{52a}$$

and

$$\operatorname{var}(\mathbf{p}'\mathbf{b} - a - \mathbf{L}'\mathbf{Y}) \text{ is a minimum.} \tag{52b}$$

**Theorem 2** The optimum estimator that satisfies (52) takes the form

$$\begin{aligned} \mathbf{p}'\hat{\mathbf{b}}^{(b)} &= \mathbf{p}'\theta + \mathbf{p}'\mathbf{F}\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{F}\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\theta) \\ &= \mathbf{p}'\theta + \mathbf{p}'(\mathbf{F}^{-1} + \mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\theta). \end{aligned} \tag{53}$$

*Proof.* The expectation in (52a) yields

$$a = (\mathbf{p}' - \mathbf{L}'\mathbf{X})\theta. \tag{54}$$

Employing (51a) and (51b), the quantity to be minimized is

$$v = \mathbf{p}'\mathbf{F}\mathbf{p} + \mathbf{L}'(\mathbf{X}\mathbf{F}\mathbf{X}' + \mathbf{V})\mathbf{L} - 2\mathbf{L}'\mathbf{X}\mathbf{F}\mathbf{p}. \tag{55}$$

Then, differentiating $v$ with respect to $\mathbf{L}$ and setting the result equal to zero, we obtain

$$(\mathbf{XFX}' + \mathbf{V})\mathbf{L} = \mathbf{XFp}$$

and the optimizing $\mathbf{L}$ is

$$\mathbf{L} = (\mathbf{XFX}' + \mathbf{V})^{-1}\mathbf{XFp}. \tag{56}$$

Substitution of (54) and (56) into $a + \mathbf{L}'\mathbf{Y}$ yields the first expression of (53). The equivalence of the two expressions in (53) is established by using the Woodbury (1950) matrix identity

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}, \tag{57}$$

where $\mathbf{A} = \mathbf{V}, \mathbf{B} = \mathbf{X}, \mathbf{C} = \mathbf{F}$ and $\mathbf{D} = \mathbf{X}'$ (see, for example, Gruber (1990) equation (4.1.7)).

Fill in the details in Exercise 20. ■

Substitution into (55) gives the minimum variance of (52b) as

$$\begin{aligned} v_{\min} &= \mathbf{p}'\mathbf{Fp} - \mathbf{p}'\mathbf{FX}'(\mathbf{XFX}' + \mathbf{V})^{-1}\mathbf{XFp} \\ &= \mathbf{p}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{p} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{F} + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{p}. \end{aligned} \tag{58}$$

Notice that $v_{\min}$ is less than the variance of the least-square estimator.

When $\mathbf{F} = \sigma^2\mathbf{G}^{-1}$, $\mathbf{V} = \sigma^2\mathbf{I}$ and $\theta = 0$ the estimator in (53) reduces to

$$\mathbf{p}'\hat{\mathbf{b}}^{(r)} = \mathbf{p}'(\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{X}'\mathbf{Y}, \tag{59}$$

the generalized ridge regression estimator of C. R. Rao (1975). When $\mathbf{G} = \mathrm{k}\mathbf{I}$, (59) reduces to the ridge regression estimator of Hoerl and Kennard (1970).

Ridge regression estimators are useful for data where the $\mathbf{X}'\mathbf{X}$ matrices have very small eigenvalues and some of the independent variables are highly correlated. Such data are called multicollinear. In such cases, the total variance of the least-square estimator can be very large and the estimates of the regression parameters very imprecise. One possible solution is to use ridge regression estimators instead. For more about ridge estimators, see Gruber (1998, 2010).

The estimators derived in this section from Theorem 2 are linear Bayes estimators. When the prior distribution and the population are normal, they are the same as the Bayes estimators derived using Bayes theorem and are the mean of the posterior distribution.

## 4. CONSEQUENCES OF ESTIMATION

Properties of $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and the resulting consequences are now discussed. The topics dealt with in this section are based solely on the two properties so far attributed to $\mathbf{e}$, that $E(\mathbf{e}) = \mathbf{0}$ and $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. In Section 5, we consider distributional properties of the estimators based on the normality of the **e**'s. However, that assumption is not made here. The general case of $\text{var}(\mathbf{e}) = \mathbf{V}$ is left largely to the reader.

We shall also mention a few of the interesting properties of the ridge estimator and compare and contrast them with those of the least-square estimator where appropriate.

### a. Unbiasedness

Since $\hat{\mathbf{b}}$ is the b.l.u.e. of $\mathbf{b}$ for $\mathbf{V} = \sigma^2\mathbf{I}$, it is unbiased. This can also be shown directly. We have

$$E(\hat{\mathbf{b}}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{b}. \tag{60}$$

Since the expected value of $\hat{\mathbf{b}}$ is $\mathbf{b}$, the estimator $\hat{\mathbf{b}}$ is unbiased. Of course, this implies that in $\hat{\mathbf{b}}' = [\hat{b}_0 \quad \hat{\boldsymbol{\beta}}]$ the estimator $\hat{\boldsymbol{\beta}}$ is also unbiased.

However, the ridge estimator is a biased estimator. Indeed

$$E(\hat{\mathbf{b}}^{(r)}) = (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}$$

and

$$Bias(\hat{\mathbf{b}}^{(r)}) = \mathbf{b} - (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{G}\mathbf{b}. \tag{61}$$

### b. Variances

The variance covariance matrix of $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is given by

$$\begin{aligned} \text{var}(\hat{\mathbf{b}}) &= E[\hat{\mathbf{b}} - E(\hat{\mathbf{b}})][\hat{\mathbf{b}} - E(\hat{\mathbf{b}})]' \\ &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{y} - E(\mathbf{y})][\mathbf{y}' - E(\mathbf{y}')]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}\mathbf{e}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2. \end{aligned} \tag{62}$$

A similar result holds for $\hat{\boldsymbol{\beta}}$ using the partitioned form of $(\mathbf{X}'\mathbf{X})^{-1}$ shown in (34). With $\mathbf{S} = \mathcal{X}'\mathcal{X}$ the result (62) becomes

$$\text{var}\begin{bmatrix} \hat{b}_o \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \frac{1}{N} + \bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1} \\ -(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}} & (\mathcal{X}'\mathcal{X})^{-1} \end{bmatrix}\sigma^2.$$

Hence analogous to (62)

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathcal{X}'\mathcal{X})^{-1}\sigma^2, \tag{63}$$

$$\text{var}(b_0) = \frac{\sigma^2}{N} + \bar{\mathbf{x}}'\text{var}(\hat{\boldsymbol{\beta}})\bar{\mathbf{x}} = \left[\frac{1}{N} + \bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}}\right]\sigma^2 \tag{64}$$

and

$$\operatorname{cov}(\hat{b}_0, \hat{\boldsymbol{\alpha}}) = -\bar{\mathbf{x}}' \operatorname{var}(\hat{\boldsymbol{\alpha}}) = -\bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\sigma^2. \tag{65}$$

The ridge regression estimator (59) has variance given by

$$\operatorname{var}(\hat{\mathbf{b}}^{(r)}) = (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\sigma^2. \tag{66}$$

Its variance is less than that of the least-square estimator in the sense of the Loewner ordering meaning that for all $\mathbf{p}$ $\operatorname{var}(\mathbf{p}'\hat{\mathbf{b}}^{(r)}) \leq \operatorname{var}(\mathbf{p}'\hat{\mathbf{b}})$.

Another measure of efficiency of an estimator is the matrix mean square error. Given a vector $\theta$ with estimator $\hat{\theta}$

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)(\hat{\theta} - \theta)' \\ &= E[\hat{\theta} - E(\hat{\theta})][\hat{\theta} - E(\hat{\theta})]' + E[\theta - E(\hat{\theta})][\theta - E(\hat{\theta})]' \\ &= \operatorname{var}(\hat{\theta}) + [Bias(\hat{\theta})]^2. \end{aligned} \tag{67}$$

The MSE of the least-square estimator is equal to its variance because it is unbiased. For the ridge estimator using (61), (66), and (67), we have

$$MSE(\hat{\mathbf{b}}^{(r)}) = (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}(\mathbf{G}\mathbf{b}\mathbf{b}'\mathbf{G} + \sigma^2\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}. \tag{68}$$

It can be shown that in the sense of the Loewner ordering (see Section 2.3)

$$MSE(\hat{\mathbf{b}}^{(r)}) \leq MSE(\hat{\mathbf{b}}) \text{ or equivalently for every vector } \mathbf{p}\ MSE(\mathbf{p}'\hat{\mathbf{b}}^{(r)}) \leq MSE(\mathbf{p}'\hat{\mathbf{b}})$$

if and only if

$$\mathbf{b}'(2\mathbf{G}^{-1} + (\mathbf{X}'\mathbf{X})^{-1})^{-1}\mathbf{b} \leq \sigma^2. \tag{69}$$

See Gruber (2010) for a proof.

### c. Estimating $E(y)$

The estimator $\hat{\mathbf{b}}$ can be used for estimating $E(y)$. Analogous to the model

$$E(y) = b_0 + b_1x_1 + \cdots + b_kx_k$$

we have

$$\widehat{E(y)} = \hat{b}_0 + \hat{b}_1x_1 + \cdots + \hat{b}_kx_k,$$

as was illustrated at the end of each of the examples in Section 1. If

$$\mathbf{x}_0' = \begin{bmatrix} x_{00} & x_{01} & x_{02} & \cdots & x_{0k} \end{bmatrix} \tag{70}$$

is a set of $x$ values with $x_{00} = 1$ for which we wish to estimate the corresponding value of $E(y)$ that estimator is

$$\widehat{E(y_0)} = \hat{b}_0 + \hat{b}_1 x_{01} + \cdots + \hat{b}_k x_{0k} = \mathbf{x}_0' \hat{\mathbf{b}}. \tag{71}$$

**Example 5 An Estimate of $E(y)$** In Example 2, we had

$$\widehat{E(y)} = 7.000 + 6.250x_1 - 0.625x_2$$

For $\mathbf{x}_0' = [\,1 \quad 12 \quad 32\,]$, we would have

$$\widehat{E(y_0)} = 7.000 + 6.250(12) - 0.625(32) = 62.$$

□

The result in (71) and the illustration in Example 5 above is called the *estimated expected value* of $y$ corresponding to the set of $x$ values $x_{00}, x_{01}, \ldots, x_{0k}$. When this set of $x$'s is one of those in the data, $\mathbf{x}_0'$ is a row of $\mathbf{X}$ in (18), in which case (71) is an element of $\mathbf{X}\hat{\mathbf{b}}$. Corresponding to $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$ of (17), we have

$$\widehat{E(\mathbf{y})} = \mathbf{X}\hat{\mathbf{b}}. \tag{72}$$

These are the estimated expected values of $y$ corresponding to the $N$ observed values of $y$ in the data. They are sometimes called *fitted y values*, or *estimated y values*. These names can be misleading because (71) and (72) are both estimates of expected values of $y$. They correspond in (71) to any set of predetermined $x$'s in $\mathbf{x}_0'$ of (70), and in (72) to the observed $x$'s in $\mathbf{X}$.

Variances of the estimators (71) and (72) are readily obtained using (63). Thus, for any $\mathbf{x_0}$

$$v(\widehat{E(y_0)}) = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0. \tag{73}$$

**Example 6 Computation of the Variances** Using the data of Example 2,

$$v(\widehat{E(y_0)}) = \begin{bmatrix} 1 & 12 & 32 \end{bmatrix} \frac{\boldsymbol{\sigma}^2}{2880} \begin{bmatrix} 50656 & 1840 & -1960 \\ 1840 & 400 & -160 \\ -1960 & -160 & 100 \end{bmatrix} \begin{bmatrix} 1 \\ 12 \\ 32 \end{bmatrix} = 2.2556\sigma^2.$$

For the observed $x$'s we have

$$\text{var}[\widehat{\mathbf{E}(\mathbf{y})}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\sigma}^2. \tag{74}$$

After substituting $\mathbf{X} = \begin{bmatrix} 1 & \mathbf{X}_1 \end{bmatrix}$ from (31) and using $\mathcal{X}$ from (39), equation (74) reduces to

$$\begin{aligned} \text{var}[\widehat{E(\mathbf{y})}] &= \frac{\sigma^2}{N}\mathbf{1}\mathbf{1}' + \mathcal{X}\text{var}(\hat{\boldsymbol{\alpha}})\mathcal{X}' \\ &= \frac{\sigma^2}{N}\mathbf{1}\mathbf{1}' + \mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\sigma^2. \end{aligned} \tag{75}$$

Corresponding to $\mathbf{x}_0'$, the expected $y$ value is $E(y_0)$, estimated by $\widehat{E(y_0)}$ of (71). In contrast, we consider a future observation, $y_f$, say, corresponding to some vector of $x$ values $\mathbf{x}_f$, say. Then, by the model $y_f = \mathbf{x}_f'\mathbf{b} + e_f$, where $e_f$ is a random error term that is neither observed or estimated. Hence, the best available prediction of $y_f$, which we shall call $\tilde{y}_f$, is $\tilde{y}_f = \mathbf{x}_f'\hat{\mathbf{b}}$. Thus, $\mathbf{x}_f'\hat{\mathbf{b}}$ can be used both as a prediction of a future observation corresponding to $\mathbf{x}_f'$ as well as for its more customary use, that of an estimator of the expected value $E(y_f)$ corresponding to $\mathbf{x}_f'$. The first of these uses prompts inquiring how some future observation $y_f$ varies about its prediction $\tilde{y}_f = \mathbf{x}_f'\hat{\mathbf{b}}$. To do so, we consider the deviation of any $y_f$ from $\tilde{y}_f$:

$$y_f - \tilde{y}_f = y_f - \mathbf{x}_f'\hat{\mathbf{b}} = \mathbf{x}_f'(\mathbf{b} - \hat{\mathbf{b}}) + e_f.$$

The variance of this deviation is derived by noting that, because $y_f$ is thought of as an observation obtained independently of those used in deriving $\hat{\mathbf{b}}$, we have $\hat{\mathbf{b}}$ and $e_f$ being independent and so $\text{cov}(\hat{\mathbf{b}}, e_f) = 0$. Hence,

$$v(y_f - \tilde{y}_f) = \mathbf{x}_f'v(\hat{\mathbf{b}} - \mathbf{b})\mathbf{x}_f + v(e_f) = [\mathbf{x}_f'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f + 1]\sigma^2. \tag{76}$$

Thus, the estimated value of $y$ corresponding to $\mathbf{x}_f$ is $E(y_f) = \mathbf{x}_f'\hat{\mathbf{b}}$ as in (71) with variance $\mathbf{x}_f'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f\sigma^2$ similar to (73). The predicted value of an observation corresponding to $\mathbf{x}_f$ is the same value $\mathbf{x}_f'\hat{\mathbf{b}} = \tilde{y}_f$ with the variance of the deviations of the $y$ values (corresponding to $\mathbf{x}_f$) from this prediction being $[\mathbf{x}_f'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f + 1]\sigma^2$ of (76). These results are true for any value of $\mathbf{x}_f$. The variance of $\tilde{y}_f$ is of course $\sigma^2$ at all times.

**Example 7 Prediction from a Future Observation and Its Variance** In Example 2, let $\mathbf{x}_f' = \begin{bmatrix} 1 & 15 & 40 \end{bmatrix}$. Then $\tilde{y}_f = \begin{bmatrix} 1 & 15 & 40 \end{bmatrix}\begin{bmatrix} 7.000 \\ 6.250 \\ -0.625 \end{bmatrix} = 75.75$.

Also

$$\begin{aligned} v(y_f - 75.75) &= \left[ \begin{bmatrix} 1 & 15 & 40 \end{bmatrix} \frac{1}{2880} \begin{bmatrix} 50656 & 1840 & -1960 \\ 1840 & 400 & -160 \\ -1960 & -160 & 100 \end{bmatrix} \begin{bmatrix} 1 \\ 15 \\ 40 \end{bmatrix} + 1 \right] \sigma^2. \\ &= 3.45\sigma^2 \end{aligned}$$

□

### d. Residual Error Sum of Squares

It is convenient to use the symbol $\hat{\mathbf{y}}$ for $\widehat{E(\mathbf{y})}$, the vector of estimated expected values of $y$ corresponding to the vector of observations $\mathbf{y}$. Thus, we have

$$\hat{\mathbf{y}} = \widehat{E(\mathbf{y})} = \mathbf{X}\hat{\mathbf{b}}. \tag{77}$$

The vector of deviations of the observed $y_i$'$s$ from their corresponding predicted values, the vector of *residuals*, is therefore

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}. \tag{78}$$

Observe that the matrix used in (78) is idempotent. We shall use this fact repeatedly in the sequel. Indeed.

$$\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{79}$$

is symmetric and idempotent and

$$[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X} = \mathbf{0}. \tag{80}$$

**Example 8 A Residual** In Example 2, we had $y_3 = 51, x_{13} = 10, x_{23} = 32$. We get that

$$\hat{y}_3 = 7 + 6.250(10) - (0.625)(32) = 49.5.$$

Then the residual is

$$y_3 - \hat{y}_3 = 51 - 49.5 = 1.5. \qquad \square$$

The sum of squares of the deviations of the observed $y_i$'s from their estimated expected values is usually known as the *residual error sum of squares*, combining the traditional name "error" with "residual," which is perhaps more appropriately descriptive in view of the definition of $e_i$ given in (3). The symbol SSE is used. It is given by

$$\text{SSE} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}). \tag{81}$$

Computing procedures for the SSE are derived from substituting (78) into (81) and using (79) and (80). This gives

$$\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} \tag{82}$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} \tag{83}$$

because $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. This is a convenient form for computing SSE. The term $\mathbf{y}'\mathbf{y}$ in (83) is the total sum of squares of the observations. The term $\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ is the sum of

the products of the elements of the solution $\hat{\mathbf{b}}$ with their corresponding elements of the right-hand side, $\mathbf{X}'\mathbf{y}$, of the equations from which $\hat{\mathbf{b}}$ is derived, namely $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$.

Note, however, that in so describing (83), these right-hand side elements must be exactly as they are in the normal equations. Thus, if, when solving $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$, some or all of the normal equations are amended by factorizing out some common factors, then it is not the right-hand sides of the equations so amended that are used in $\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ of (83) but the $\mathbf{X}'\mathbf{y}$ of the original normal equations.

An expression for SSE involving $\hat{\boldsymbol{\theta}}$ and $\mathcal{X}'\mathbf{y}$ can also be established. Let $\tilde{y} = y - \bar{y}$. Equation (83) is equivalent to

$$\begin{aligned} \text{SSE} &= \mathbf{y}'\mathbf{y} - [\bar{y} - \hat{\boldsymbol{\theta}}'\bar{\mathbf{x}} \quad \tilde{\mathbf{b}}'] \begin{bmatrix} N\bar{y} \\ \mathbf{X}_1'\mathbf{y} \end{bmatrix} \\ &= \mathbf{y}'\tilde{\mathbf{y}} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'(\mathbf{X}_1'\mathbf{y} - N\bar{y}\bar{\mathbf{x}}) \qquad (84) \\ &= \mathcal{y}'\mathcal{y} - \hat{\boldsymbol{\theta}}\mathcal{X}'\mathbf{y}. \qquad (85) \end{aligned}$$

The term $\mathcal{y}'\mathcal{y}$ denotes the corrected sum of squares of the $y$'s. The form of (85) is completely analogous to that of (83) and it is equally, if not more, useful for computing purposes. We have that $\mathcal{y}'\mathcal{y}$ is the corrected sum of squares of the $y$'s and $\hat{\boldsymbol{\theta}}\mathcal{X}'\mathbf{y}$ is the sum of products of elements of the solution of the corresponding elements of the right-hand side, $\mathcal{X}'\mathbf{y}$, of the equations from which $\hat{\boldsymbol{\theta}}$ is derived, namely $\mathcal{X}'\mathcal{X}\hat{\boldsymbol{\theta}} = \mathcal{X}'\mathbf{y}$. Observe that $\mathcal{X}'\mathbf{y} = \mathcal{X}'\,\tilde{\mathcal{y}}$ because $\mathcal{X}'\,\mathbf{y} = \mathcal{X}'(\mathbf{y} - \bar{y}\mathbf{1}_N)$ and by (33) and (39) $\mathcal{X}'\mathbf{1}_N = 0$.

### e. Estimating the Residual Error Variance

In (82), SSE is written as a quadratic form in **y:**

$$\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$$

Therefore, with $\mathbf{y}$ having mean $\mathbf{X}\mathbf{b}$ and variance $\mathbf{I}\sigma^2$, the expected value of SSE is from Theorem 4 of Chapter 2.

$$\begin{aligned} E[SSE] &= \text{tr}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{I}\sigma^2 + \mathbf{b}'\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X}\mathbf{b} \\ &= r[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2 \\ &= [N - r(\mathbf{X})]\sigma^2, \end{aligned}$$

making use of (79) and (80) and the fact that the trace of an idempotent matrix equals its rank. Hence, an unbiased estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N - r(\mathbf{X})} = \frac{\text{SSE}}{N - r} \qquad (86)$$

using $r$ for $r(\mathbf{X})$, the rank of $\mathbf{X}$. We use $r$ even though we know that in this full-rank situation

$$r = r\,(\text{X}) = k + 1$$

This is to emphasize that it is the rank of $\mathbf{X}$ and not just the number of $x$ variables plus one. It also makes for an easier transition to the non-full-rank case, to be studied in Chapter 5, where it is essential to use $r(\mathbf{X})$.

### f. Partitioning the Total Sum of Squares

The total sum of squares, which we shall call SST, is

$$\text{SST} = \mathbf{y}'\mathbf{y} = \sum_{i=1}^{N} y_i^2$$

The sum of squares of deviations of observed $y_i$'s from their predicted values is

$$\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$$

as in (83) and (84). The difference

$$\text{SSR} = \text{SST} - \text{SSE} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = N\bar{y}^2 + \hat{\boldsymbol{\theta}}'\mathcal{X}'\ \mathbf{y}$$

represents that portion of SST that is attributed to having fitted the regression, and so it is called the *sum of squares due to regression*, SSR. It is also often called the *reduction in sum of squares.* The partitioning of SST can be summarized in a manner that serves as a foundation for developing the traditional analysis of variance (ANOVA) table.

$$\begin{array}{l} \text{SSR} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = N\bar{y}^2 + \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} \\ \text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} \\ \hline \text{SST} = \mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{y}. \end{array} \tag{87}$$

Now suppose the model had no $x$ variables in it but had simply been $y_i = b_0 + e_i$.Then $\hat{b}_0$ would be $\bar{y}$ and SSR would become $N\bar{y}^2$. This we recognize as the usual correction for the mean which is written

$$\text{SSR} = N\bar{y}^2$$

Then in (87) we see that

$$\text{SSR} = \text{SSM} + \hat{\boldsymbol{\theta}}'\ \mathcal{X}'\mathbf{y}$$

and so we can call

$$\text{SSR}_\text{m} = \text{SSR} - \text{SSM} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{X}\hat{\boldsymbol{\theta}}$$

the regression sum of squares corrected for the mean. In this way, (87) becomes

$$\begin{array}{r} \text{SSM} = N\bar{y}^2 \\ \text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathcal{X}\hat{\boldsymbol{\theta}} \\ \text{SSE} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} \\ \hline \text{SST} = \mathbf{y}'\mathbf{y} \end{array} \tag{88}$$

Similar to $\text{SSR}_\text{m}$, we have

$$\text{SST}_\text{m} = \text{SST} - \text{SSM} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 = \boldsymbol{y}'\boldsymbol{y} \tag{89}$$

With (89), $\text{SSR}_\text{m}$ and SSE of (88) can be summarized as

$$\begin{array}{r} \text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} \\ \text{SSE} = \boldsymbol{y}'\boldsymbol{y} - \hat{\mathbf{b}}'\mathcal{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\mathbf{b}}'\mathcal{X}'\mathbf{y} \\ \hline \text{SST}_\text{m} = \boldsymbol{y}'\boldsymbol{y} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 \end{array} \tag{90}$$

This format is identical to that of (87). In one, (87), uncorrected sum of squares are used with total SST, and in the other, (90), corrected sums of squares are used with total $\text{SST}_\text{m}$. The error terms are the same in the two cases, namely SSE.

The summary shown in (90) is the basis of the traditional analysis of variance table for fitting linear regression. Distributional properties of these sums of squares are considered in Section 5.

### g. Multiple Correlation

The multiple correlation coefficient is a measure of the goodness of fit of a regression line. It is estimated as the product moment correlation between the observed $y_i$'s and the predicted $\hat{y}_i$'s. It is denoted by $R$ and can be calculated as $R^2 = \frac{\text{SSR}}{\text{SST}}$ for the no-intercept model and as

$$R^2 = \frac{\text{SSR}_\text{m}}{\text{SST}_\text{m}} \text{ for the intercept model.} \tag{91}$$

This we now show.

In the no-intercept model, we ignore the mean $\bar{y}$. The product moment correlation between the $y_i$'s and the $\hat{y}_i$'s is defined by

$$R^2 = \frac{\sum_{i=1}^{N} y_i\hat{y}_i}{\left(\sum_{i=1}^{N} y_i^2\right)\left(\sum_{i=1}^{N} \hat{y}_i^2\right)} = \frac{(\mathbf{y}'\hat{\mathbf{y}})^2}{(\mathbf{y}'\mathbf{y})(\hat{\mathbf{y}}'\hat{\mathbf{y}})} \tag{92}$$

With $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, it can be shown (see Exercise 9) that (92) reduces to $R^2 = \text{SSR/SST}$ as in (91).

For the intercept model, the definition of $R^2$ is

$$R^2 = \frac{\left[\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]}{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2} \tag{93}$$

To simplify this expression, we use

$$\bar{y} = \frac{\mathbf{1}_N'\mathbf{y}}{N} \quad \text{and} \quad \mathbf{1}_N'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{1}_N'. \tag{94}$$

The second equation in (94) holds true because

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} 1_N' \\ \mathbf{X}_1' \end{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} 1_N'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{X}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} = \mathbf{X}' = \begin{bmatrix} 1_N' \\ \mathbf{X}_1' \end{bmatrix}.$$

These results together with (89) lead (see Exercise 9) to (93) reducing to $R^2 = \text{SSR}_\text{m}^2/\text{SST}_\text{m}(\text{SSR}_\text{m}) = \text{SSR}_\text{m}/\text{SST}_\text{m}$ as in (91).

Intuitively, the ratio SSR/SST or $\text{SSR}_\text{m}/\text{SST}_\text{m}$ has appeal because it represents that fraction of the total sum of squares which is accounted for by fitting the regression model. Thus, although $R$ has traditionally been thought about and used as a multiple correlation coefficient in some sense, its more frequent use nowadays is in the form of $R^2$ where it represents the fraction of the total sum of squares accounted for by fitting the model.

Care must be taken in using these formulae for $R^2$, for, although $\text{SSR}_\text{m}$ and $\text{SST}_\text{m}$ have been defined in the intercept model as $\text{SSR}-N\bar{y}^2$ and $\text{SST}-N\bar{y}^2$, the value of SSR used in the intercept model is not the same as its value in the corresponding no-intercept model resulting in different values of $R^2$. This will be illustrated in Example 9 which follows.

**Example 9 Computation of Predicted Values, Their Variances, Residuals, and Sums of Squares** In (28), we found for the data of Example 2 that the least-square estimates were

$$\hat{\mathbf{b}} == \begin{bmatrix} 7.000 \\ 6.250 \\ -0.625 \end{bmatrix}.$$

Then the vector

$$\widehat{E(\mathbf{y})} = \hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \begin{bmatrix} 1 & 6 & 28 \\ 1 & 12 & 40 \\ 1 & 10 & 32 \\ 1 & 8 & 36 \\ 1 & 9 & 34 \end{bmatrix} \begin{bmatrix} 7.000 \\ 6.250 \\ -0.625 \end{bmatrix} = \begin{bmatrix} 27 \\ 57 \\ 49.5 \\ 34.5 \\ 42 \end{bmatrix}.$$

Hence from (74) using $(\mathbf{X}'\mathbf{X})^{-1}$ of (26)

$$\begin{aligned} \operatorname{var}(\hat{\mathbf{y}}) &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 \\ &= \begin{bmatrix} 1 & 6 & 28 \\ 1 & 12 & 40 \\ 1 & 10 & 32 \\ 1 & 8 & 36 \\ 1 & 9 & 34 \end{bmatrix} \frac{\sigma^2}{2880} \begin{bmatrix} 50656 & 1840 & -1960 \\ 1840 & 400 & -160 \\ -1960 & -160 & 100 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6 & 12 & 10 & 8 & 9 \\ 28 & 40 & 32 & 36 & 34 \end{bmatrix}. \end{aligned}$$

From (75), using $\mathcal{X}$ of (43) and $(\mathcal{X}'\mathcal{X})^{-1}$ of (44),

$$\begin{aligned} \operatorname{var}(\hat{\mathbf{y}}) &= \frac{1}{5}\mathbf{1}\mathbf{1}'\sigma^2 + \mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\boldsymbol{\sigma}^2 \\ &= \frac{1}{5}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \boldsymbol{\sigma}^2 \\ &\quad + \begin{bmatrix} -3 & -6 \\ 3 & 6 \\ 1 & -2 \\ -1 & 2 \\ 0 & 0 \end{bmatrix} \frac{\boldsymbol{\sigma}^2}{144} \begin{bmatrix} 20 & -8 \\ -8 & 5 \end{bmatrix} \begin{bmatrix} -3 & 3 & 1 & -1 & 0 \\ -6 & 6 & -2 & 2 & 0 \end{bmatrix}. \end{aligned}$$

After carrying out the arithmetic, it will be found that both forms reduce to

$$\operatorname{var}(\hat{\mathbf{y}}) = \begin{bmatrix} 0.7 & -0.3 & 0.2 & 0.2 & 0.2 \\ -0.3 & 0.7 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.7 & -0.3 & 0.2 \\ 0.2 & 0.2 & -0.3 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix} \boldsymbol{\sigma}^2.$$

An estimate of this is obtained by replacing $\boldsymbol{\sigma}^2$ by $\hat{\boldsymbol{\sigma}}^2$ as will be shown below.

From $\mathbf{y}$ and $\hat{\mathbf{y}}$, we obtain the vector of residuals

$$(\mathbf{y} - \hat{\mathbf{y}}) = \begin{bmatrix} 30 \\ 60 \\ 51 \\ 36 \\ 33 \end{bmatrix} - \begin{bmatrix} 27 \\ 57 \\ 49.5 \\ 34.5 \\ 42 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 1.5 \\ 1.5 \\ -9 \end{bmatrix}.$$

Recall that SSE is the sum of the squares of the residuals or $(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$. Hence,

$$\text{SSE} = 3^2 + 3^2 + (1.5)^2 + (1.5)^2 + (-9)^2 = 103.5.$$

The alternative form of SSE, given in (83) is

$$\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}.$$

With $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{5} y_i^2 = 9486$ in the data of Example 2, $\hat{\mathbf{b}}$ from (28), and $\mathbf{X}'\mathbf{y}$ from (27), we obtain

$$\begin{aligned} \text{SSE} &= 9486 - (\tfrac{1}{24}) \begin{bmatrix} 168 & 150 & -15 \end{bmatrix} \begin{bmatrix} 210 \\ 1995 \\ 7290 \end{bmatrix} \\ &= 9486 - 9382.5 = 103.5 \text{ as before.} \end{aligned}$$

Likewise, using the form given in (85),

$$\begin{aligned} \text{SSE} &= 9486 - 5(42^2) - \tfrac{1}{24} \begin{bmatrix} 150 & -15 \end{bmatrix} \begin{bmatrix} 105 \\ 150 \end{bmatrix} \\ &= 9486 - 8820 - 562.5 = 103.5 \text{ again.} \end{aligned}$$

Hence, in (86), our estimate of $\sigma^2$ becomes

$$\hat{\sigma}^2 = \frac{103.5}{(5-3)} = 51.75.$$

From the calculations for SSE, the summaries in (87), (88), and (90) are as shown in Table 3.1. From the last of these, $R^2$ is $\text{SSR}_\text{m}/\text{SST}_\text{m} = 562.5/666 = 0.845$, since the model being used is the intercept model. If a no-intercept model were to be used on these data, the formal expression for $R^2$ would be SSR/SST, although not with the value of SSR shown in Table 3.1, because that is the value of SSR for the intercept model. For the no-intercept model, these data normal equations for $\hat{\mathbf{b}}$ and $\mathbf{X}'\mathbf{y}$ are given in (29). Thus,

$$\text{SSR} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \begin{bmatrix} 5.9957 & -0.3542 \end{bmatrix} \begin{bmatrix} 1995 \\ 7290 \end{bmatrix} = 9379.3.$$

**TABLE 3.1 Partitioning of Sum of Squares: Intercept Model**

| Eqs.(87) | Eqs.(89) | Eqs.(90) |
|---|---|---|
| | SSM = 8820 | |
| SSR = 9382.5 | $SSR_m$ = 562.5 | $SSR_m$ = 562.5 |
| SSE = 103.5 | SSE = 103.5 | SSE = 103.5 |
| SST = 9486 | SST = 9486 | $SST_m$ = 666 |

This value of SSR is different from that obtained for the intercept model from (87) given in Table 3.1. The corresponding value of $R^2$ is 9379.3/9486 = 0.989.

Notice that the two $R^2$ are different. The intercept model accounts for 84.5% of the variation while the no-intercept model accounts for 98.9% of the variation.We can use the software package R to obtain residuals and fitted values.

```
> income=c(30,60,51,36,33)
> years=c(6,12,10,8,9)
> age=c(28,40,32,36,34)
> lm.r=lm(income~years+age)
```

We give the least-square coefficients, the residuals and the fitted values using R. The output follows.

```
>coef(lm.r)
(Intercept)       years          age
      7.000       6.250      -0.625
> resid(lm.r)
  1    2    3    4    5
  3.0  3.0  1.5  1.5 -9.0
> fitted(lm.r)
   1    2    3    4    5
27.0 57.0 49.5 34.5 42.0
```

A plot of the residuals vs. the fitted values is given in Figure 3.2.

## 5. DISTRIBUTIONAL PROPERTIES

Up to now, we made no assumptions about the distribution of **e**. We will assume that **e** is normally distributed in order to develop confidence intervals and tests of hypothesis about the regression parameters. In what follows, we assume that

$$\mathbf{e} \sim N(0, \sigma^2\mathbf{I}).$$

This will enable us to derive the distributions of $\mathbf{y}$, $\hat{\mathbf{b}}, \hat{\sigma}^2$, and the various sums of squares using the results of Chapter 2.

### a. The Vector of Observations y is Normal

From $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ we have $\mathbf{y} - \mathbf{Xb} = \mathbf{e}$ and therefore,

$$\mathbf{y} \sim N(\mathbf{Xb}, \sigma^2\mathbf{I}_N).$$

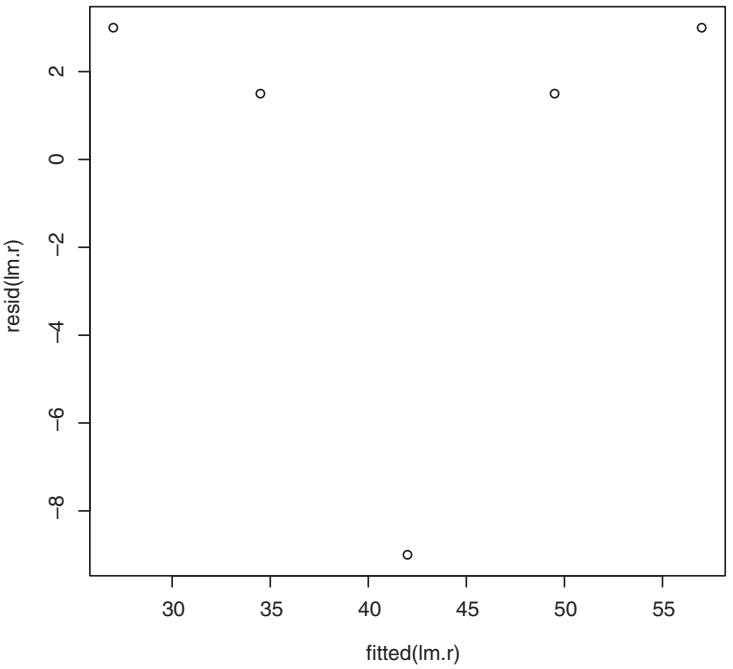



**FIGURE 3.2** Plot of Residuals vs. Fitted Values

### b. The Least-square Estimator $\hat{\mathbf{b}}$ is Normal

The least-square estimator $\hat{\mathbf{b}}$ is normally distributed because it is a linear function of $\mathbf{y}$. The mean and variance were derived in (60) and (62). Thus,

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N(\mathbf{b}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2).$$

Using the same reasoning, we have that $\hat{\hat{\mathbf{b}}}$ is normally distributed. From (60) and (63)

$$\hat{\boldsymbol{\beta}} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}\mathbf{y} \sim N(\boldsymbol{\beta}, (\mathcal{X}'\mathcal{X})^{-1}\sigma^2)$$

### c. The Least-square Estimator $\hat{\mathbf{b}}$ and the Estimator of the Variance $\hat{\sigma}^2$ are Independent

We have

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and

$$\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

However, by (80),

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = 0.$$

The independence of $\hat{\mathbf{b}}$ and $\hat{\sigma}^2$ follows from Theorem 6 of Chapter 2.

### d. The Distribution of SSE/$\sigma^2$ is a $\chi^2$ Distribution

From (82), SSE is a quadratic form in $\mathbf{y}$ with matrix $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Then $SSE/\sigma^2 = \mathbf{y}'(1/\sigma^2)\mathbf{P}\mathbf{y}$. By (79), $\mathbf{P}$ is idempotent and var($\mathbf{y}$) = $\sigma^2\mathbf{I}$. Therefore, $(1/\sigma^2)\mathbf{P}\sigma^2\mathbf{I}$ is idempotent. From Theorem 5 of Chapter 2,

$$\text{SSE}/\sigma^2 \sim \chi^{2\prime}[\mathbf{r}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'], \mathbf{b}'\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X}\mathbf{b}/2\sigma^2] \qquad (95)$$

From (79) and (80), (95) reduces to $\text{SSE}/\sigma^2 \sim \chi^2_{N-r}$, where $r = r(\mathbf{X})$. Thus,

$$(N - r)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-r}.$$

### e. Non-central $\chi^{2\prime}$ s

We have shown that SSE/$\sigma^2$ has a central $\chi^2$-distribution. We will now show that SSR, SSM, and $\text{SSR}_\text{m}$ have non-central $\chi^2$-distributions. Furthermore, these terms are independent of SSE. Thus we are led to $F$-statistics that have non-central $F$-distributions. These in turn are central $F$-distributions under certain null hypothesis. Tests of these hypotheses are established.

From (87), we have

$$\text{SSR} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is idempotent and its product with $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the null matrix. Applying Theorem 7 of Chapter 2, SSR and SSE are independent. By Theorem 5 of the same chapter,

$$\text{SSR}/\sigma^2 \sim \chi^{2\prime}\{r[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}], \mathbf{b}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2\} = \chi^{2\prime}(r, \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2.$$

Similarly, in (88),

$$\text{SSM} = N\bar{y}^2 = \mathbf{y}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{y},$$

where $N^{-1}\mathbf{1}\mathbf{1}'$ is idempotent and its product with $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the null matrix. Therefore, SSM is distributed independently of SSE and

$$\text{SSM}/\sigma^2 \sim \chi^{2\prime}[r(N^{-1}\mathbf{1}\mathbf{1}'), \mathbf{b}'\mathbf{X}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{X}\mathbf{b}/2\sigma^2] = \chi^{2\prime}[1, (1\mathbf{X}\mathbf{b})^2/2N\sigma^2].$$

Also in (90),

$$\text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathcal{X}'\hat{\boldsymbol{\theta}}.$$

Since $\hat{\mathbf{b}} \sim N[\mathbf{b}, (\mathcal{X}'\mathcal{X})^{-1}]\sigma^2$,

$$\text{SSR}_\text{m}/\sigma^2 \sim \chi^{2\prime}[r(\mathcal{X}'\mathcal{X}), \boldsymbol{\theta}\mathcal{X}'\mathcal{X}\boldsymbol{\theta}/2\sigma^2] = \chi^{2\prime}(r - 1, \boldsymbol{\theta}\mathcal{X}'\mathcal{X}\boldsymbol{\theta}/2\sigma^2).$$

Furthermore, $\text{SSR}_\text{m}$ can be expressed as $\mathbf{y'Qy}$ where $\mathbf{Q} = \mathcal{X}(\mathcal{X}'\mathcal{X}')^{-1}\mathcal{X}'$ where $\mathbf{Q}$ is idempotent and its products with $\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'$ and $N^{-1}\mathbf{11}'$ are the null matrix. By Theorem 7 of Chapter 2, $\text{SSR}_\text{m}$ is independent of both SSE and SSM.

Finally, of course

$$\mathbf{y}'\mathbf{y}/\sigma^2 \sim \chi^{2'}(N, \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2).$$

### f. *F*-distributions

Applying the definition of the non-central $F$-distribution to the foregoing results, it follows that the $F$-statistic

$$F(R) = \frac{\text{SSR}/r}{\text{SSE}/(N-r)} \sim F'(r, N-r, \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2). \tag{96}$$

Similarly

$$F(M) = \frac{\text{SSM}/1}{\text{SSE}/(N-r)} \sim F'[1, N-r, (\mathbf{1}'\mathbf{X}\mathbf{b})^2/2N\sigma^2] \tag{97}$$

and

$$F(R_\text{m}) = \frac{\text{SSR}_\text{m}/(r-1)}{\text{SSE}/(N-r)} \sim F'[r-1, N-r, \boldsymbol{\mathscr{b}}\mathcal{X}'\mathcal{X}\boldsymbol{\mathscr{b}}/2\sigma^2]. \tag{98}$$

Under certain null hypotheses, the non-centrality parameters in (96)–(98) are zero and these non-central $F$'s then become central $F$'s and thus provide us with statistics to test these hypotheses. This is discussed subsequently.

### g. Analyses of Variance

Calculation of the above $F$-statistics can be summarized in analyses of variance 4s. An outline of such tables is given in (87), (88), and (90). For example, (87) and the calculation of (96) are summarized in Table 3.2.

**TABLE 3.2 Analysis of Variance for Fitting Regression**

| Source of Variation | d.f.[a] | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Regression | $r$ | $\text{SSR} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ | $\text{MSR} = \dfrac{\text{SSR}}{r}$ | $F(R) = \dfrac{\text{MSR}}{\text{MSE}}$ |
| Residual error | $N-r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| Total | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | | |

[a] $r = r(\mathbf{X}) = k + 1$ when there are K regression variable ($x$'s).

**TABLE 3.3 Analysis of Variance, Showing a Term for the Mean**

| Source of Variation[a] | d.f.[b] | Sum of Squares | Mean Square | $F$-Statistics |
|---|---|---|---|---|
| Mean | 1 | $\text{SSM} = N\bar{y}^2$ | $\text{MSM} = \text{SSM}/1$ | $F(M) = \dfrac{\text{MSM}}{\text{MSE}}$ |
| Regression (c.f.m.) | $r-1$ | $\text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$ | $\text{MSR}_\text{m} = \dfrac{\text{SSR}_\text{m}}{r-1}$ | $F(R_\text{m}) = \dfrac{\text{MSR}_\text{m}}{\text{MSE}}$ |
| Residual error | $N-r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | | |

[a] c.f.m. = corrected for the mean.
[b] $r = r(\mathbf{X}) = k + 1$ when there are K regression variables ($x$'s).

This table summarizes not only the sums of squares-already summarized in (87) but also degrees of freedom (d.f.) of the associated $\chi^2$-distributions. In the mean squares, the sum of squares divided by the degrees of freedom, the table also shows calculation of the numerator and the denominator of $F$. It also shows the calculation of $F$ itself. Thus, the analysis of variance table is simply a convenient summary of the steps involved in calculating the $F$-statistic.

In a manner similar to Table 3.2, (88) and the $F$-ratios of (97) and (98) are summarized in Table 3.3.

The abbreviated form of this, based on (90) and showing only the calculation of (98) is as shown in Table 3.4.

Tables 3.2, 3.3, and 3.4 are all summarizing the same thing. They show development of the customary form of this analysis, namely 3.4. Although it is the form customarily seen, it is not necessarily the most informative. Table 3.3 has more information because it shows how SSR of table 3.2 is partitioned into SSM and $\text{SSR}_\text{m}$ the regression sum of squares corrected for the mean (c.f.m.). Table 3.4 is simply an abbreviated version of Table 3.3 with SSM removed from the body of the table and subtracted from SST to give $\text{SST}_\text{m} = \text{SST} - \text{SSM} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$, the corrected sum of squares of the $y$ observations. Thus, although Table 3.4 does not show $F(M) = \text{MSM/MSE}$, it is identical to Table 3.3 insofar as $F(R_\text{m}) = \text{MSR}_\text{m}/\text{MSE}$ is concerned.

**TABLE 3.4 Analysis of Variance (Corrected for the Mean)**

| Source of Variation[a] | d.f.[b] | Sum of Squares | Mean Square | $F$-Statistics |
|---|---|---|---|---|
| Regression (c.f.m) | $r-1$ | $\text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$ | $\text{MSR}_\text{m} = \dfrac{\text{SSR}_\text{m}}{r-1}$ | $F(R_\text{m}) = \dfrac{\text{MSR}_\text{m}}{\text{MSE}}$ |
| Residual error | $N-r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - N\bar{y}^2 - \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| Total | $N-1$ | $\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$ | | |

[a] c.f.m. = corrected for the mean.
[b] $r = r(\mathbf{X}) = k + 1$ when there are K regression variables.

### h. Tests of Hypotheses

Immediately after (96)–(98), we made the comment that those results provide us with statistics for testing hypothesis. We illustrate this now. In Section 6, we will take up the general linear hypothesis.

In Table 3.2, the statistic $F(R)$, as shown in (96), is distributed as a non-central $F$ with non-centrality parameter $\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2$. The non-centrality parameter is zero under the null hypothesis $H_0$: $\mathbf{b} = \mathbf{0}$. In this case, $F(R)$ has a central $F$-distribution $F_{r,N-r}$. The statistic $F(R)$ may be compared to the tabulated values to test the hypothesis. We may specify a level $\alpha$, any level we want. In statistical practice, popular $\alpha$ levels are 0.10, 0.05, and 0.01.

When $F(R) \geq$ tabulated $F_{r,N-r}$ at the $100\alpha\%$ level, we reject the null hypothesis $H_0$: $\mathbf{b} = \mathbf{0}$ at that level of significance. Otherwise, we fail to reject $H_0$. We may find the tabulated value from tables, using a handheld calculator, the Texas Instrument TI 84, for example, or a statistical software package. We may also calculate the $p$-value which is the lowest level of significance where the null hypothesis is rejected by finding the probability that $F_{r,N-r}$ is greater than the calculated statistic $F(R)$ and reject $H_0$ at level $\alpha$ when $\alpha > p$-value. To find the $p$-value, we need either a statistical handheld calculator (TI 83 or 84, for example) or statistical software package.

Apropos assuming the model $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$, we might then say, borrowing a phrase from Williams (1959), that when $F(R)$ is significant, there is "concordance of the data with this assumption" of the model. That means that the model accounts for a significant portion of the variation. This does not mean that this model for the particular set of $x$'s is necessarily the most suitable model. Indeed, there may be a subset of those $x$'s that are as significant as the whole. There may be further $x$'s which when used alone or in combination with some or all of the $x$'s already used that are significantly better than those already used. Furthermore, there may be nonlinear functions of those $x$'s that are at least or more suitable than using linear functions of the $x$'s. None of these contingencies is inconsistent with $F(R)$ being significant and the ensuing conclusion that the data are in concordance with the model $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$.

In addition to what was discussed in the previous paragraph, a statistically significant model might only account for a small percentage of the variation. To judge the suitability of a model, other facts must be taken into consideration besides statistical significance.

The non-centrality parameter of the $F$-statistic $F(M)$ of Table 3.3 is, as in (97), $(\mathbf{1}'\mathbf{X}\mathbf{b})^2/2N\sigma^2$. For the numerator of this expression

$$\mathbf{1}'\mathbf{X}\mathbf{b} = \mathbf{1}'E(\mathbf{y}) = E(\mathbf{1}'\mathbf{y}) = E(N\bar{y}) = NE(\bar{y}).$$

Hence the non-centrality parameter in (97) is $N[E(\bar{y})]^2/2\sigma^2$, which is zero under the hypothesis $H_0$: $E(\bar{y}) = 0$. The statistic $F(M)$ is distributed as $F_{1,N-r}$. Thus, it can be used to test $H_0$ meaning that it can be used to test the hypothesis that the expected value of the mean of the observed values is zero. This is an interpretation of the

phrase "testing the mean" sometimes used for describing the test based on $F(M)$. Equivalently, $\sqrt{F(M)}$ has the $t$-distribution with $N - r$ degrees of freedom because

$$F(M) = \frac{N\bar{y}^2}{\hat{\sigma}^2} = \left[\frac{\bar{y}}{\hat{\sigma}/\sqrt{N}}\right]^2$$

is the square of a $t$ random variable.

Another way of looking at the test provided by $F(M)$ is based on the model $E(y_i) = b_0$. The reduction sum of squares for fitting this model is SSM, and the non-centrality parameter in (97) is then $Nb_0^2/2\sigma^2$. Hence $F(M)$ can be used to test whether the model $E(y_i) = b_0$ accounts for variation in the $y$ variable.

In using a test based on $F(R)$, we are testing the hypothesis that all the $b_i$'s including $b_0$, are simultaneously zero. However, for the null hypothesis $H_0 : \tilde{\mathbf{b}} = \mathbf{0}$, that is, that just the $b$'s corresponding to the $x$ variables are zero, then the test is based on $F(R_{\mathrm{m}})$ in Tables 3.3 and 3.4. This is so because, from (98), we see that the non-centrality parameter in the non-central $F$-distribution of $F(R_{\mathrm{m}})$ is zero under the null hypothesis $H_0 : \boldsymbol{\mathscr{C}} = \mathbf{0}$. In this case, $F(R_{\mathrm{m}})$ has a central $F$-distribution on $r - 1$ and $N - r$ degrees of freedom. Thus $F(R_{\mathrm{m}})$ provides a test of hypothesis $H_0 : \tilde{\mathbf{b}} = \mathbf{0}$. If $F(R_{\mathrm{m}})$ is significant, the hypothesis is rejected. This is not to be taken as evidence that all the elements of $\tilde{\mathbf{b}}$ are non-zero. It simply means that at least one of them may be. If $F(M)$ has first been found significant, then $F(R_{\mathrm{m}})$ being significant indicates that a model with the $x$'s explains significantly more of the variance in the $y$ variable than does the model $E(y_i) = b_0$.

Tests using $F(M)$ and $F(R_{\mathrm{m}})$ are based on the numerators SSM and $\mathrm{SSR}_{\mathrm{m}}$. As shown earlier in Section 5e, these random variables are statistically independent. The $F$'s themselves are not independent because they have the same denominator mean square. The probability of rejecting at least one of the hypotheses $b_0 = 0, \tilde{\mathbf{b}} = 0$ each at level $\alpha$ would be between $\alpha$ and $2\alpha$. One way to insure a simultaneous test of both hypotheses $F(M)$ and $F(R_{\mathrm{m}})$ did not have a significance level greater than $\alpha$ would be to perform each individual test at level $\alpha/2$. This is an example of a multiple comparisons procedure. For more information on this important topic, see, for example, Miller (1981).

Another possibility is the case where $F(M)$ is not significant but $F(R_{\mathrm{m}})$ is.[1] This would be evidence that even though $E(\bar{y})$ might be zero, fitting the $x$'s does explain variation in the $y$ variable. An example of a situation where this might occur is when the $y$ variable can have both positive and negative values, such as weight gains in beef cattle, where in fact some gains may be in fact losses, that is, negative gains.

**Example 10 Analysis of Variance Results for Regression in Example 2** The results are presented in Table 3.5 below. Using the summaries shown in Table 3.1, the analyses of variance in Tables 3.2–3.4 are shown in Table 3.5. The first part of Table 3.5 shows $F(R) = 60.4$, with 3 and 2 degrees of freedom. Since the tabulated value of the $F_{3,2}$-distribution is 19.15 at $\alpha = .05$ and $F(R) = 60.5 > 19.15$, we conclude that the model accounts for a significant (at the 5% level) portion of the variation of

[1] S.R. Searle is grateful to N.S.Urquhart for emphasizing this possibility.

**TABLE 3.5 Tables 3.2, 3.3, and 3.4 for Data of Example 2**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| | Table 3.2 | | | |
| Regression | 3 | SSR = 9382.5 | 3127.5 | $F(R) = 3073.5/51.75 = 60.43$ |
| Residual error | 2 | SSE = 103.5 | 51.75 | |
| Total | 5 | SST = 9486 | | |
| | Table 3.3 | | | |
| Mean | 1 | SSM = 8820 | 8820 | $F(M) = 8820/51.75 = 170.4$ |
| Regression c.f.m. | 2 | $SSR_m = 562.5$ | 281.25 | $F(R_m) = 281.25/51.75 = 5.4$ |
| Residual error | 2 | SSE = 103.5 | 51.75 | |
| Total | 5 | SST = 9486 | | |
| | Table 3.4 | | | |
| Regression c.f.m. | 2 | $SSR_m = 562.5$ | 281.25 | $F(R_m) = 281.25/51.75 = 5.4$ |
| Residual error | 2 | SSE = 103.5 | 51.75 | |
| Total | 4 | $SST_m = 666$ | | |

the $y$ variable. However, since the $p$-value is 0.016, the model does not account for a significant portion of the $y$ variable at the 1% level. Likewise, $F(M)$ of the Table 3.3 portion of Table 3.5 has 1 and 2 degrees of freedom and since $F(M) = 170.4 > 18.51$, the tabulated value of the $F_{1,2}$-distribution at the 5% level, we reject the hypothesis that $E(\bar{y}) = 0$. In this case, the $p$-value is 0.0059. We would also reject the hypothesis $E(\bar{y}) = 0$ at $\alpha = .01$. The value of $F_{1,2}$ for the 1% level of significance is 98.5. Finally, since $F(R_m) = 5.4 < 19.0$, the tabulated value of the $F_{2,2}$-distribution, we fail to reject the hypothesis that $b_1 = b_2 = 0$. The $p$-value in this case is $0.16 > 0.05$. This test provides evidence that the $x$'s are contributing little in terms of accounting for variation in the $y$ variable. Most of the variation is accounted for by the mean, as is evident from the sums of squares values in the Table 3.3 section of Table 3.5. As is true generally, the Table 3.4 section is simply an abbreviated form of the Table 3.3 section, omitting the line for the mean. Just how much of the total sum of squares has been accounted for by the mean is, of course, not evident in the Table 3.4 section. This is a disadvantage to Table 3.4, even though its usage is traditional. □

### i. Confidence Intervals

We shall now develop confidence intervals for regression coefficients. On the basis of normality assumptions discussed in Section 5b, we know that $\hat{\mathbf{b}}$ has a normal distribution. As a result

$$\frac{\hat{b}_i - b_i}{\sqrt{a^{ii}\sigma^2}} \sim N(0, 1), \tag{99}$$

for $i = 0,1,2,\ldots,$ or $k$ where in accord with the development of (63) and (64)

$$a^{00} = \frac{1}{N} + \bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}} \tag{100}$$

and for $i = 1, 2, \ldots, k$

$$a^{ii} = i\text{th diagonal element of } (\mathcal{X}'\mathcal{X})^{-1}. \tag{101}$$

With these values of $a^{ii}$, and in (99) replacing $\sigma^2$ by $\hat{\sigma}^2$ of (86), we have

$$\frac{\hat{b}_i - b_i}{\sqrt{a^{ii}\hat{\sigma}^2}} \sim t_{N-r}, \tag{102}$$

where $t_{N-r}$ represents the $t$-distribution on $N - r$ degrees of freedom.

Define $t_{N-r,\alpha,L}$ and $t_{N-r,\alpha,U}$ as a pair of lower and upper limits respectively of the $t_{N\text{-}r}$-distribution such that

$$\Pr\left\{t \le t_{N-r,\alpha,L}\right\} + \Pr\left\{t \ge t_{N-r,\alpha,U}\right\} = \alpha$$

As a result, we have

$$\Pr\left\{t_{N-r,\alpha,L} \le t \le t_{N-r,\alpha,U}\right\} = 1 - \alpha$$

for $t \sim t_{N\text{-}r.}$ Then by (102),

$$\Pr\left\{t_{N-r,\alpha,L} \le \frac{\hat{b}_i - b_i}{\sqrt{a^{ii}\hat{\sigma}^2}} \le t_{N-r,\alpha,U}\right\} = 1 - \alpha. \tag{103}$$

Rearrangement of the probability statement in the form

$$\Pr\left\{\hat{b}_i - \hat{\sigma}t_{N-r,\alpha,U}\sqrt{a^{ii}} \le b_i \le \hat{b}_i - \hat{\sigma}t_{N-r,\alpha,L}\sqrt{a^{ii}}\right\} = 1 - \alpha$$

provides

$$\left(\hat{b}_i - \hat{\sigma}t_{N-r,\alpha,U}\sqrt{a^{ii}}, \hat{b}_i - \hat{\sigma}t_{N-r,\alpha,L}\sqrt{a^{ii}}\right) \tag{104}$$

as a $100(1 - \alpha)\%$ confidence interval. Usually symmetric confidence intervals are utilized. For this confidence interval to be symmetric with respect to $b_i$, we need

$$-t_{N-r,\alpha,L} = t_{N-r,\alpha,U} = t_{N-r,\frac{1}{2}\alpha}, \text{ where } \Pr\left\{t \ge t_{N-r,\frac{1}{2}\alpha}\right\} = \frac{1}{2}\alpha \tag{105}$$

and the interval (104) becomes

$$\hat{b}_i \pm \hat{\sigma}t_{N-r,\frac{1}{2}\alpha}\sqrt{a^{ii}}, \tag{106}$$

of width $2\hat{\sigma}t_{N-r,\frac{1}{2}\alpha}\sqrt{a^{ii}}$.

When the degrees of freedom are large ($N - r > 100$ say) the distribution in (102) is approximately $N(0, 1)$. Define $\nu_{\alpha,L}$ and $\nu_{\alpha,U}$ such that

$$\Pr\{\nu_{\alpha,L} \leq \nu \leq \nu_{\alpha,U}\} = 1 - \alpha, \text{ for } \nu \sim N(0, 1). \tag{107}$$

The values $\nu_{\alpha,L}$ and $\nu_{\alpha,U}$ can be used in (104) in place of $t_{N-r,\alpha,L}$ and $t_{N-r,\alpha,U}$. In particular, for a symmetric confidence interval,

$$\nu_{\alpha,L} = -\nu_{\alpha,U} = z_{\frac{1}{2}\alpha}, \text{ where } (2\pi)^{-\frac{1}{2}} \int_{z_{\frac{1}{2}\alpha}}^{\infty} e^{-\frac{1}{2}x^2} dx = \frac{1}{2}\alpha.$$

The resulting confidence interval is

$$\hat{b}_i \pm \hat{\sigma} z_{\frac{1}{2}\alpha} \sqrt{a^{ii}}. \tag{108}$$

Tabulated values of $z_{\frac{1}{2}\alpha}$ for a variety of values of $\frac{1}{2}\alpha$ are available in Table I of the statistical tables online.

Confidence intervals for any linear combination of the $b$'s, $\mathbf{q}'\mathbf{b}$, say, can be established in a like manner. The argument is unchanged, except that at all stages $b_i$ and $\hat{b}_i$ are replaced by $\mathbf{q}'\mathbf{b}$ and $\mathbf{q}'\hat{\mathbf{b}}$, respectively, and $a^{ii}\hat{\sigma}^2$ is replaced by $\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}\hat{\sigma}^2$. Thus, from (106) and (108), the symmetric confidence interval for $\mathbf{q}'\mathbf{b}$ is

$$\mathbf{q}'\hat{\mathbf{b}} \pm \hat{\sigma} t_{N-r,\frac{1}{2}\alpha} \sqrt{\mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}} \tag{109}$$

with $z_{\frac{1}{2}\alpha}$ replacing $t_{N-r,\frac{1}{2}\alpha}$ when $N - r$ is large.

In equation (71), we developed $\mathbf{x}_0'\hat{\mathbf{b}}$ as the estimator of $E(y_0)$ corresponding to a set of $x$'s in $\mathbf{x}_0'$. Result (109) now provides a confidence interval on $\mathbf{x}_0'\mathbf{b}$, namely

$$\mathbf{x}'_0\hat{\mathbf{b}} \pm \hat{\sigma} t_{N-r,\frac{1}{2}\alpha} \sqrt{\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}. \tag{110}$$

In the case of simple regression involving only one $x$ variable (where $k = 1$ and $r = 2$ as in the footnote to Table 3.4), $\mathbf{x}_0' = \begin{bmatrix} 1 & x_0 \end{bmatrix}$ and (110) becomes

$$\begin{bmatrix} 1 & x_0 \end{bmatrix} \begin{bmatrix} \bar{y} - \hat{b}\bar{x} \\ \hat{b} \end{bmatrix} \pm \hat{\sigma} t_{N-2,\frac{1}{2}\alpha} \sqrt{\begin{bmatrix} 1 & x_0 \end{bmatrix} \begin{bmatrix} N & N\bar{x} \\ N\bar{x} & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} 1 \\ x_0 \end{bmatrix}}.$$

This simplifies to

$$\bar{y} + b(x_0 - \bar{x}) \pm \hat{\sigma} t_{N-2,\frac{1}{2}\alpha} \sqrt{\frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}}, \tag{111}$$

the familiar expression for the confidence interval on $E(y)$ in a simple regression model (see, for example, p. 170 of Steel and Torrie (1960)). Plotting the values of this interval for a series of values of $x_0$ provides the customary confidence belt for the regression line

$$y = b_0 + bx.$$

**Example 11 Confidence Intervals for Predicted Values in Example 1**

| | Fit | lwr | upr |
|---|---|---|---|
| 1 | 26.25 | 0.9675721 | 51.53243 |
| 2 | 57.75 | 32.4675721 | 83.03243 |
| 3 | 47.25 | 25.2444711 | 69.25553 |
| 4 | 36.75 | 14.7444711 | 58.75553 |
| 5 | 42.00 | 20.4390731 | 63.56093 |

$x$ = years of schooling beyond sixth grade, $y$ = income □

Confidence bands for the predicted values in Example 1 are plotted in Figure 3.3. A confidence interval for an estimated observation is called a *tolerance or a prediction interval.* In keeping with the variance given in (76), the prediction interval comes from using $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 + 1$ instead of $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ in (110). In line with (110), it reduces for simple regression to

$$\bar{y} + b(x_0 - \bar{x}) \pm \hat{\sigma} t_{N-2,\frac{1}{2}\alpha} \sqrt{1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}}. \tag{112}$$

## j. More Examples

First, we give an example of a non-symmetric confidence interval.

**Example 12 A Non-symmetric 95% Confidence Interval for $b_1$** The non-symmetric interval will be calculated using (104) for a regression coefficient from the data of Example 2. We have the point estimates $b_1 = 6.250$, from (28) $\hat{\sigma} = 7.20$

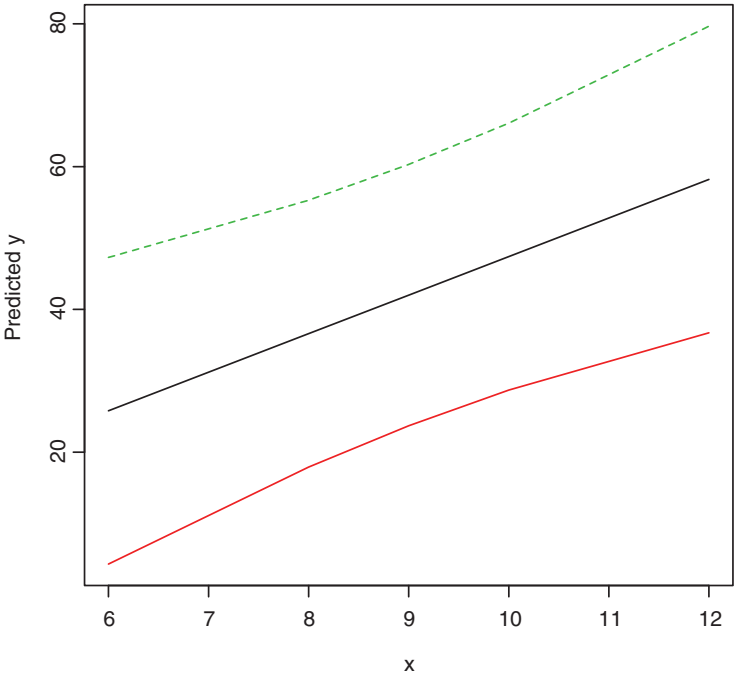



**FIGURE 3.3** Confidence Band for Predicted Values in Example 2

and $N - r = 2$ from Table 3.5. From (101) and (44) we have $a^{11} = 20/144 = 0.139$. Then in (104), a non-symmetric confidence band for $b_1$ is given by

$$(6.25 - 7.20t_{2,\alpha,U}\sqrt{0.139}, 6.25 - 7.20t_{2,\alpha,L}\sqrt{0.139})$$

or

$$(2.08 - 2.68t_{2,\alpha,U}, 2.08 - 2.68t_{2,\alpha,L}).$$

We shall set this confidence interval up so that the left-hand tail of the $t_2$-distribution has probability 0.04 and the right-hand tail has probability 0.01. We could use any tail probabilities that add up to 0.05 and get a different confidence interval for each case. Using a TI 83 or TI 84 calculator, we need to find $L$ and $U$ so that

$$P(t \geq t_{2,.05,L}) = 0.96 \quad \text{and} \quad P(\text{t} \geq t_{2,.05,U}) = 0.01.$$

We get that $t_{2,.05,L} = -3.32$ and $t_{2,.05,U} = 6.96$. Then the confidence interval becomes

$$(6.25 - 2.68(6.96),\ (6.25 - 2.68(-3.32)),\ \text{or}\ (-12.4, 15.15). \tag{113}$$

Of course it is questionable that there would be a situation that would lead to a non-symmetric confidence interval with the $t$-distribution. However, Example 11 illustrates how such intervals may be calculated and emphasizes the fact that there are many such intervals because there are many values $t_{N-r,\alpha,L}$ and $t_{N-r,\alpha,U}$ that satisfy (103). There is only one symmetric confidence interval.

**Example 13 A Symmetric 95% Confidence Interval** From Table 2 (see web page) or calculator $t_{2,.975} = -t_{2,.025} = -4.30$. Hence, our confidence interval will be $6.25 \pm 2.68(4.30) = (-5.274, 17.774)$. This confidence interval contains zero so if we were to test the hypothesis $H_0$:$b_1 = 0$ versus the two-sided alternative hypothesis $H_1 : b_1 \neq 0$, we would fail to reject $H_0$. □

**Example 14 A Simultaneous 95% Confidence Interval on $b_1$ and $b_2$** We need to find 97.5% confidence intervals on $b_1$ and $b_2$ to guarantee that the confidence level of the simultaneous confidence interval is at most .05. To do this, we need $t_{.0125} = 6.205$. The confidence intervals are for $b_1$

$$6.25 \pm 6.205(2.68) \text{ or } (-10.38,22.88)$$

and for $b_2$

$$-0.2083 \pm 6.205(.45) \text{ or } (-3,2.58).$$

Since both confidence intervals contain zero, a test of $H_0$: $b_1 = 0$, $b_2 = 0$ versus an alternative that at least one of the regression coefficients was not zero would fail to reject $H_0$ at a significance level of at most 0.05. □

**Example 15 Summary of Results for Example 2 using R**

```
Call:
lm(formula = income ~ years + age)

Residuals:
   1    2    3    4    5
 3.0  3.0  1.5  1.5 -9.0



Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.000     30.170   0.232    0.838
years          6.250      2.681   2.331    0.145
age           -0.625      1.341  -0.466    0.687
```

Confidence intervals on predicted values

```
predict(lm.r,level=0.95,interval="confidence")
   fit       lwr      upr
1 27.0  1.103535 52.89647
2 57.0 31.103535 82.89647
3 49.5 23.603535 75.39647
4 34.5  8.603535 60.39647
5 42.0 28.157757 55.84224
>
```

□

### k. Pure Error

Data sometimes have the characteristic that the set of $x$'s corresponding to several $y$'s are the same. For example, in the case of an experiment with data taken from Chen et al. (1991) in Montgomery, Runger and Hubele (2007), the relationship between noise exposure and hypertension was investigated. The independent variable $x$ represented sound pressure level in decibels while the dependent variable $y$ represented blood pressure rise in millimeters of mercury. The data were as follows:

| $y$ | 1 | 0 | 1 | 2 | 5 | 1 | 4 | 6 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 60 | 63 | 65 | 70 | 70 | 70 | 80 | 90 | 80 | 80 |

| $y$ | 5 | 4 | 6 | 8 | 4 | 5 | 7 | 9 | 7 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 85 | 89 | 90 | 90 | 90 | 90 | 94 | 100 | 100 | 100 |

It was reproduced with the kind permission of John Wiley & Sons.

Observe that there are three $x$ observations of 70, 3 of 80, 4 of 90, and 3 of 100 and one each of 60, 63, 85, 89, and 95. These are called *repeated* $x$'s. Their presence provides a partitioning of SSE into two terms, one that represents "lack of fit" of the model and the other that represents "pure error." Description is given in terms of simple regression involving one $x$ variable. Extension to several $x$'s is straightforward.

Suppose $x_1, x_2, \ldots, x_p$ are the $p$ distinct values of the $x$'s where $x_i$ occurs in the data $n_i$ times, that is, with $n_i$ $y$ values $y_{ij}$ for $j = 1, 2, \ldots, n_i$ and for $i = 1, 2, \ldots, p$. For all $i$, $n_i \geq 1$, we will write

$$n_{\cdot} = \sum_{i=1}^{p} n_i = N.$$

Then

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}^2 - \hat{\boldsymbol{\theta}}' \mathbf{X}' \mathbf{y} \\ &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}_{\cdot\cdot}^2 - \hat{\boldsymbol{\theta}}' \left( \sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij} y_{ij} - N\bar{x}_{\cdot\cdot}\bar{y}_{\cdot\cdot} \right), \end{aligned}$$

with $N - r$ degrees of freedom can be partitioned into $\text{SSPE} = \sum_{i=1}^{p} \left[ \sum_{i=1}^{n_i} y_{ij}^2 - n_i(\bar{y}_{i.})^2 \right]$ with $N - p$ degrees of freedom and

$$\text{SSLF} = \text{SSE} - \text{SSPE with } p - r \text{ degrees of freedom.}$$

In this form, SSE/$(N-p)$, known as the mean square due to *pure error*, is an estimator of $\sigma^2$. The mean square SSLF/$(p-r)$ is due to *lack of fit* of the model. It provides a test of the lack of fit by comparing

$$F(\text{LF}) = \frac{\text{SSLF}/(p-r)}{\text{SSPE}/(p-r)}$$

against $F_{p-2,N-p}$. Significance indicates that the model is inadequate. Lack of significance as Draper and Smith (1998) pointed out means that there is no reason to doubt the adequacy of the model. In this case, SSE/$(N-2)$ provides a pooled estimator of $\sigma^2$.

**Example 16 SAS Output Illustrating Lack of Fit Test for Sound Data**

The SAS System

The GLM Procedure

**Class Level Information**

| **Class** | **Levels** | **Values** |
|---|---|---|
| **lackofit** | 10 | 60 63 65 70 80 85 89 90 94 100 |

**Number of Observations Read** 20

**Number of Observations Used** 20

The SAS System

The GLM Procedure

Dependent Variable:y

| **Source** | **DF** | **Sum Squares** | **Mean Square** | **F Value** | Pr > **F** |
|---|---|---|---|---|---|
| **Model** | 9 | 100.0666667 | 11.1185185 | 4.61 | 0.0128 |
| **Error** | 10 | 24.1333333 | 2.41333333 | | |
| **Corrected Total** | 19 | 124.200000 | | | |

| **R-square** | **Coeff Var** | **Root MSE** | **y Mean** |
|---|---|---|---|
| 0.805690 | 36.12769 | 1.552491 | 4.300000 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | Pr > **F** |
|---|---|---|---|---|---|
| *x* | 1 | 92.93352510 | 92.93352510 | 38.51 | 0.0001 |
| **lackofit** | 8 | 7.13314156 | 0.89164270 | 0.37 | 0.9142 |

The slope of the regression equation is highly significant. However, there is no significant lack of fit, so there is no reason to doubt the adequacy of the model.

A simple SAS program that would generate the above output and some additional information is

```
data sound;
input y  x  lackofit;
datalines;
1 60  60
.
.
.
proc glm;
class lackofit;
model y=x lackofit;
run;
```

## 6. THE GENERAL LINEAR HYPOTHESIS

### a. Testing Linear Hypothesis

The literature of linear models abounds with discussions of different kinds of hypotheses that can be of interest in widely different fields of application. Four hypothesis of particular interest are:

(i) $H$: $\mathbf{b} = \mathbf{0}$, the hypothesis that all of the elements of $\mathbf{b}$ are zero;
(ii) $H$: $\mathbf{b} = \mathbf{b_0}$, the hypothesis that $b_i = b_{i0}$ for $i = 1, 2, \ldots, k$, that is, that each $b_i$ is equal to some specified value $b_{i0}$;
(iii) $H$: $\lambda'\mathbf{b} = m$, that some linear combination of the elements of $\mathbf{b}$ equals a specified constant;
(iv) $H$: $\mathbf{b_q} = \mathbf{0}$, that some of $b_i$'s, $q$ of them where $q < k$ are zero.

We show that *all* of the linear hypothesis above and others are special cases of a general procedure even though the calculation of the $F$-statistics may appear to differ from one hypothesis to another.

The general hypothesis we consider is

$$\text{H: } \mathbf{K'b} = \mathbf{m},$$

where $\mathbf{b}$, of course, is the $(k + 1)$-order vector of parameters of the model, $\mathbf{K'}$ is any matrix of $s$ rows and $k + 1$ columns and $\mathbf{m}$ is a vector of order $s$ of specified constants. There is only one limitation on $\mathbf{K'}$; it must have full row rank, that is, $r(\mathbf{K'}) = s$. This simply means that the linear functions of $\mathbf{b}$ must be linearly independent. The hypothesis being tested must be made up of linearly independent functions of $\mathbf{b}$ and must contain no functions that are linear functions of others therein. This is quite reasonable because it means, for example, that if the hypothesis relates to $b_1 - b_2$ and

$b_2 - b_3$, then there is no point in having it relate explicitly to $b_1 - b_3$. This condition on $\mathbf{K}'$ is not at all restrictive in limiting the application of the hypothesis $H$: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ to real problems. It is not necessary to require that $\mathbf{m}$ be such that the system of equations $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is consistent because this is guaranteed by $\mathbf{K}'$ being of full rank.

We now develop the $F$-statistic to test the hypothesis $H$: $\mathbf{K}'\mathbf{b} = \mathbf{m}$. We know that

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}), \hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \hat{\mathbf{b}} \sim N[\mathbf{b}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2].$$

Therefore,

$$\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m} \sim N[\mathbf{K}'\mathbf{b} - \mathbf{m}, \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\sigma^2]$$

By virtue of Theorem 5 in Chapter 2, the quadratic form

$$Q = (\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})$$

in $(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})$ with matrix $[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}$ has a non-central $\chi^2$-distribution. We have that

$$\frac{Q}{\sigma^2} \sim \chi^{2'}\left\{s, \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{2\sigma^2}\right\}. \tag{114}$$

We now show the independence of $Q$ and SSE using Theorem 7 of Chapter 2. We first express $Q$ and SSE as quadratic forms of the same normally distributed random variable. We note that the inverse of $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ used in (114) exists because $\mathbf{K}'$ has full row rank and $\mathbf{X}'\mathbf{X}$ is symmetric. In equation (114), we replace $\hat{\mathbf{b}}$ with $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Then equation (114) for $Q$ becomes

$$Q = [\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{m}]'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{m}].$$

The matrix $\mathbf{K}'$ has full-column rank. By the corollary to Lemma 5 in Section 3 of Chapter 2, $\mathbf{K}'\mathbf{K}$ is positive definite. Thus $(\mathbf{K}'\mathbf{K})^{-1}$ exists. Therefore,

$$\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{m} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}].$$

As a result, $Q$ may be written

$$Q = [\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$[\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}].$$

The next step is to get the quadratic form for SSE into a similar form as $Q$. Recall that

$$\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

Since $\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}] = \mathbf{0}$ and $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]\mathbf{X} = \mathbf{0}$, we may write

$$\text{SSE} = [\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}].$$

We have expressed both $Q$ and SSE as quadratic forms in the normally distributed vector $\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}$. Also the matrices for $Q$ and SSE are both idempotent, so we again verify that they have $\chi^{2'}$-distributions. More importantly, the product of the matrices for $Q$ and SSE are null. We have that

$$[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0}.$$

Therefore by Theorem 7 of Chapter 2, $Q$ and SSE are distributed independently. This gives us the $F$-distribution needed to test the hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$. We have that

$$\begin{aligned} F(H) &= \frac{Q/s}{\text{SSE}/[N - r(\mathbf{X})]} = \frac{Q}{s\hat{\sigma}^2} \\ &\sim F'\left\{ s, N - r(\mathbf{X}), \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{2\sigma^2} \right\}. \end{aligned} \tag{115}$$

Under the null hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ $F(H) \sim F_{s,N-r(\mathbf{X})}$. Hence, $F(H)$ provides a test of the null hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ and the $F$- statistic for testing this hypothesis is

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})}{s\hat{\sigma}^2} \tag{116}$$

with $s$ and $N - r$ degrees of freedom.

The generality of this result merits emphasis. It applies for any linear hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$. The only limitation is that $\mathbf{K}'$ has full-row rank. Other than this, $F(H)$ can be used to test any linear hypothesis whatever. No matter what the hypothesis is, it only has to be written in the form $\mathbf{K}'\mathbf{b} = \mathbf{m}$. Then, $F(H)$ of (116) provides the test. Having once solved the normal equations for the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ and so obtained $(\mathbf{X}'\mathbf{X})^{-1}$, $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2$, the testing of $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ can be achieved by immediate application of $F(H)$. The appeal of this result is illustrated in Section 6c for the four hypothesis listed at the beginning of this section. Notice that $\hat{\sigma}^2$ is universal to every application of $F(H)$. Thus, in considering different hypotheses, the only term that changes is $Q/s$.

### b. Estimation Under the Null Hypothesis

A natural question to ask when considering the null hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ is "What is the estimator of $\mathbf{b}$ under the null hypothesis?" This might be especially pertinent following non-rejection of the hypothesis by the preceding $F$ test. The desired estimator, $\hat{\mathbf{b}}_c$, say, is readily obtainable using constrained least squares.

Thus, $\hat{\mathbf{b}}_c$ is derived so as to minimize $(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)$ subject to the constraint $\mathbf{K}'\mathbf{b} = \mathbf{m}$.

With $2\theta'$ as a vector of Lagrange multipliers, we minimize

$$L = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c) + 2\theta'(\mathbf{K}'\hat{\mathbf{b}}_c - \mathbf{m})$$

with respect to the elements of $\hat{\mathbf{b}}_c$ and $\theta'$. Differentiation with respect to these elements leads to the equations

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\mathbf{b}}_c + \mathbf{K}\theta &= \mathbf{X}'\mathbf{y} \\ \mathbf{K}'\hat{\mathbf{b}}_c &= \mathbf{m}. \end{aligned} \tag{117a}$$

The equations in (117) are solved as follows. From the first,

$$\hat{\mathbf{b}}_c = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{K}\theta) = \hat{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\theta. \tag{117b}$$

Substitution of this result into the second equation yields

$$\mathbf{K}'\hat{\mathbf{b}}_c = \mathbf{K}'\hat{\mathbf{b}} - \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\theta = \mathbf{m}.$$

Hence,

$$\theta = [\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m}). \tag{117c}$$

Substitution of $\theta$ in (117c) into (117a) gives the constrained least-square estimator

$$\hat{\mathbf{b}}_c = \hat{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m}). \tag{118}$$

The expression obtained in (118) and the $F$-statistic derived in (116) apply directly to $\hat{\boldsymbol{\alpha}}$ when the hypothesis is $\mathbf{L}'\boldsymbol{\alpha} = \mathbf{m}$ (see Exercise 12).

We have estimated $\mathbf{b}$ under the null hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$. We now show that the corresponding residual sum of squares is SSE + $Q$ where $Q$ is the numerator sum of squares of the $F$- statistic used in testing the hypothesis in (116), $F(H)$. We consider the residual $(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)$, add and subtract $\mathbf{X}\hat{\mathbf{b}}$ and show that we get SSE + $Q$.

Thus

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c)'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_c) &= [\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)]'[\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)] \\ &= (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) + (\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)'\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \\ &\quad + (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)' + (\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_c) \\ &= (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) + (\hat{\mathbf{b}} - \hat{\mathbf{b}}_c)'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}_c). \end{aligned} \tag{119}$$

The two middle terms in (119) are zero because $\mathbf{x}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{0}$.

Substituting the constrained least-square estimator (118) into (119), we get

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_c)'(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_c)\\&=\text{SSE}+(\mathbf{K}'\hat{\mathbf{b}}-\mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\&\quad\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}}-\mathbf{m})\\&=\text{SSE}+(\mathbf{K}'\hat{\mathbf{b}}-\mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}}-\mathbf{m})\\&=\text{SSE}+Q\end{aligned}\tag{120}$$

from (114).

In deriving the constrained least-square estimator, we used an exact constraint $\mathbf{K}'\mathbf{b}=\mathbf{m}$. We could have used a stochastic constraint of the form $\mathbf{m}=\mathbf{K}'\mathbf{b}+\boldsymbol{\tau}$ where $\boldsymbol{\tau}$ is a random variable and have obtained a different estimator. We shall derive these estimators and see why they are interesting in Section 6e.

### c. Four Common Hypotheses

In this section, we illustrate the expressions for $F(H)$ and $\hat{\mathbf{b}}_c$ for four commonly occurring hypotheses. We derive the $F$-statistic as special cases of that in (116).

(i) First consider $H$: $\mathbf{b}=\mathbf{0}$. The test of this hypothesis has already been considered in the analysis of variance tables. However, it illustrates the reduction of $F(H)$ to the $F$-statistic of the analysis of variance tables. To apply $F(H)$ we need to specify $\mathbf{K}'$ and $\mathbf{m}$ for the equation $\mathbf{K}'\mathbf{b}=\mathbf{m}$. We have that $\mathbf{K}'=\mathbf{I}, s=k+1$ and $\mathbf{m}=\mathbf{0}$. Thus, $[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}$ becomes $\mathbf{X}'\mathbf{X}$. Then, as before,

$$F(H)=\frac{\hat{\mathbf{b}}\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}}{(k+1)\hat{\sigma}^2}=\frac{\text{SSR}}{r}\cdot\frac{N-r}{\text{SSE}}.$$

Under the null hypothesis $F(R)\sim F_{r,\,N-r}$, where $r=k+1$.

Of course, the corresponding value of $\hat{\mathbf{b}}_c$ is

$$\hat{\mathbf{b}}_c=\hat{\mathbf{b}}-(\mathbf{X}'\mathbf{X})^{-1}[(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\hat{\mathbf{b}}=\mathbf{0}.$$

(ii) We now consider $H$: $\mathbf{b}=\mathbf{b_0}$, that is, $b_i=b_{i0}$ for all $i$. Rewriting $\mathbf{b}=\mathbf{b_0}$ as $\mathbf{K}'\mathbf{b}=\mathbf{m}$ gives
$\mathbf{K}'=\mathbf{I}, s=k+1, \mathbf{m}=\mathbf{b}_0$ and $[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}=\mathbf{X}'\mathbf{X}$. Thus,

$$F(H)=\frac{(\hat{\mathbf{b}}-\mathbf{b}_0)'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}}-\mathbf{b}_0)}{(k+1)\hat{\sigma}^2}.\tag{121}$$

An alternative expression for the numerator of (121) may be obtained. Observe that

$$\begin{aligned}(\hat{\mathbf{b}}-\mathbf{b}_0)'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}}-\mathbf{b}_0)&=(\mathbf{y}-\mathbf{X}\mathbf{b}_0)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}_0)\\&=(\mathbf{y}-\mathbf{X}\mathbf{b}_0)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}_0).\end{aligned}$$

However, the form shown in (121) is probably most suitable for computing purposes. Under the null hypothesis $F(H) \sim F_{r,N\text{-}r}$ where $r = k + 1$.

In this case, the estimator of $\mathbf{b}$ under the null hypothesis is

$$\hat{\mathbf{b}}_{\mathbf{c}} = \hat{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1}[(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\hat{\mathbf{b}} - \mathbf{b}_0) = \mathbf{b}_0.$$

(iii) Now, consider $H$: $\boldsymbol{\lambda}'\mathbf{b} = m$. In this case, we have $\mathbf{K}' = \boldsymbol{\lambda}'$, $s = 1$ and $\mathbf{m} = m$. Since $\boldsymbol{\lambda}'$ is a vector ,

$$F(H) = \frac{(\boldsymbol{\lambda}'\hat{\mathbf{b}} - m)'[\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}]^{-1}(\boldsymbol{\lambda}'\hat{\mathbf{b}} - m)}{\hat{\sigma}^2} = \frac{(\boldsymbol{\lambda}'\hat{\mathbf{b}} - m)^2}{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}\hat{\sigma}^2}.$$

Under the null hypothesis, $F(H)$ has the $F_{1,N\text{-}r}$-distribution. Hence,

$$\sqrt{F(H)} = \frac{\boldsymbol{\lambda}'\mathbf{b} - m}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}}} \sim t_{N-r}.$$

This is as one would expect because $\boldsymbol{\lambda}'\hat{\mathbf{b}}$ is normally distributed with variance $\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}$.

For this hypothesis, the value of $\hat{\mathbf{b}}_c$ is

$$\begin{aligned}\hat{\mathbf{b}}_c &= \hat{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}[\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}]^{-1}(\boldsymbol{\lambda}'\hat{\mathbf{b}} - m)\\ &= \hat{\mathbf{b}} - \left\{\frac{\boldsymbol{\lambda}'\hat{\mathbf{b}} - m}{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}}\right\}(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}.\end{aligned}$$

Observe that

$$\begin{aligned}\boldsymbol{\lambda}'\hat{\mathbf{b}}_c &= \boldsymbol{\lambda}'\hat{\mathbf{b}} - \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}[\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}]^{-1}(\boldsymbol{\lambda}'\hat{\mathbf{b}} - m)\\ &= \boldsymbol{\lambda}'\hat{\mathbf{b}} - (\boldsymbol{\lambda}'\hat{\mathbf{b}} - m) = m.\end{aligned}$$

Thus, $\hat{\mathbf{b}}_c$ satisfies the null hypothesis $H$: $\boldsymbol{\lambda}'\mathbf{b} = m$.

At this point, it is appropriate to comment on the lack of emphasis being given to the $t$-test in hypothesis testing. The equivalence of $t$-statistics with $F$-statistics with one degree of freedom in the numerator makes it unnecessary to consider $t$-tests. Whenever a $t$-test might be proposed, the hypothesis to be tested can be put in the form $H$: $\lambda'\mathbf{b} = m$ and the $F$-statistic $F(H)$ derived as here. If the $t$-statistic is insisted upon, it is then obtained as $\sqrt{F(H)}$. No further discussion of using the $t$-test is therefore necessary.

(iv) We now consider the case where the null hypothesis is that the first $q$ coordinates of $\mathbf{b}$ is zero, that is, $H : \mathbf{b_q} = \mathbf{0}$, i.e., $b_i = 0$ *for* $i = 0, 1, 2, \ldots q - 1$, *for* $q < k$. In this case, we have $\mathbf{K}' = \begin{bmatrix}\mathbf{I}_q & \mathbf{0}\end{bmatrix}$ and $\mathbf{m} = \mathbf{0}$ so that $s = q$. We write

$$\mathbf{b'_q} = \begin{bmatrix}b_0 & b_1 & \cdots & b_{q-1}\end{bmatrix}$$

and partition $\mathbf{b}$, $\hat{\mathbf{b}}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ accordingly. Thus,

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_{\mathrm{q}} \\ \mathbf{b}_{\mathrm{p}} \end{bmatrix}, \hat{\mathbf{b}} = \begin{bmatrix} \hat{\mathbf{b}}_{\mathrm{q}} \\ \hat{\mathbf{b}}_{\mathrm{p}} \end{bmatrix} \text{ and } (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{T}_{qq} & \mathbf{T}_{qp} \\ \mathbf{T}_{pq} & \mathbf{T}_{pp} \end{bmatrix},$$

where $p + q =$ the order of $\mathbf{b} = k + 1$. Then in $F(H)$ of (116)

$$\mathbf{K}'\hat{\mathbf{b}} = \hat{\mathbf{b}}_{\mathbf{q}}$$

and

$$[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1} = \mathbf{T}_{qq}^{-1},$$

giving

$$F(H) = \frac{\hat{\mathbf{b}}_q'\mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q}{q\hat{\sigma}^2}. \tag{122}$$

In the numerator, we recognize the result (e.g., Section 9.11 of Searle (1966)) of "invert part of the inverse". That means, take the inverse of $\mathbf{X}'\mathbf{X}$ and invert that part of it that corresponds to $\mathbf{b}_q$ of the hypothesis $H : \mathbf{b}_{\mathrm{q}} = \mathbf{0}$. Although demonstrated here that for a $\mathbf{b}_{\boldsymbol{q}}$ that consists of the first $q$ $b$'s in $\mathbf{b}$, it clearly applies to any subset of $q$ $b$'s. In particular, for just one $b$, it leads to the usual $F$-test on one degree of freedom, equivalent to a $t$-test (see Exercise 17).

The estimator of $\mathbf{b}$ under this hypothesis is

$$\begin{aligned} \hat{\mathbf{b}}_c &= \hat{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} \mathbf{I}_q \\ \mathbf{0} \end{bmatrix} \mathbf{T}_{qq}^{-1}(\hat{\mathbf{b}}_q - \mathbf{0}) \\ &= \hat{\mathbf{b}} - \begin{bmatrix} \mathbf{T}_{qq} \\ \mathbf{T}_{pq} \end{bmatrix} \mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q = \begin{bmatrix} \hat{\mathbf{b}}_q \\ \hat{\mathbf{b}}_p \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{b}}_q \\ \mathbf{T}_{pq}\mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{b}}_p - \mathbf{T}_{pq}\mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q \end{bmatrix}. \end{aligned}$$

Thus, the estimators of the $b$'s not in the hypothesis are $\hat{\mathbf{b}}_p - \mathbf{T}_{pq}\mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q$.

The expressions obtained for $F(H)$ and $\hat{\mathbf{b}}_c$ for these four hypotheses concerning $\mathbf{b}$ are in terms of $\hat{\mathbf{b}}$. They also apply to similar hypotheses in terms of $\hat{\tilde{\mathbf{b}}}$ (see Exercise 14), as do analogous results for any hypothesis $\mathbf{L}'\boldsymbol{\beta} = m$ (see Exercise 12).

### d. Reduced Models

We now consider, in turn, the effect of the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ of the hypotheses $\mathbf{K}'\mathbf{b} = \mathbf{m}, \mathbf{K}'\mathbf{b} = \mathbf{0},$ and $\mathbf{b}_q = 0$.

(*i*) ***The Hypothesis*** $\mathbf{K}'\mathbf{b} = \mathbf{m}$ In estimating $\mathbf{b}$ subject to $\mathbf{K}'\mathbf{b} = \mathbf{m}$, it could be said that we are dealing with a model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ on which has been imposed the limitation $\mathbf{K}'\mathbf{b} = \mathbf{m}$. We refer to the model that we start with, $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, without the limitation as the *full model.* The model with the limitation imposed $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ with $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is called the *reduced model.* For example, if the full model is

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i$$

and the hypothesis is $H : b_1 = b_2$, the reduced model is

$$y_i = b_0 + b_1(x_{i1} + x_{i2}) + b_3 x_{i3} + e_i.$$

We will now investigate the meaning of $Q$ and SSE + $Q$ in terms of sums of squares associated with the full and the reduced models. To aid description, we introduce the terms reduction(full), residual(full) and residual(reduced) for the reduction and residual after fitting the full model and the residual after fitting the reduced model. We have

$$\text{reduction(full)} = \text{SSR} \quad \text{and} \quad \text{residual(full)} = \text{SSE}.$$

Similarly,

$$\text{SSE} + Q = \text{residual(reduced)} \tag{123}$$

as established in (120). Hence,

$$Q = \text{SSE} + Q - \text{SSE} = \text{residual(reduced)} - \text{residual(full)}. \tag{124}$$

Furthermore,

$$\begin{aligned} Q &= \mathbf{y}'\mathbf{y} - \text{SSE} - [\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)] \\ &= \text{SSE} - [\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)] \\ &= \text{reduction(full)} - [\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)]. \end{aligned} \tag{125}$$

Comparison of (125) with (124) tempts one to conclude that $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is reduction(reduced), the reduction sum of squares due to fitting the reduced model. The temptation to do this is heightened by the fact that SSE + $Q$ is residual(reduced) as in (123). However, we shall show that $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is the reduction in the sum of squares due to fitting the reduced model only in special cases. It is not always so. The circumstances of these special cases are quite wide, as well as useful. First, we

show that $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$, in general, is not a sum of squares. It can be negative. To see this, observe that in

$$\begin{aligned}\mathbf{y}'\mathbf{y} - (\text{SSE} + Q) &= \text{SSR} - Q \\ &= \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - (\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbf{b}} - \mathbf{m}),\end{aligned} \tag{126}$$

the second term is a positive semi-definite quadratic form. Hence it is never negative.

If one or more of the elements of $\mathbf{m}$ is sufficiently large, that term will exceed $\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ and (126) will be negative. As a result, $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is not a sum of squares.

The reason that $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is not necessarily a reduction in the sum of squares due to fitting the reduced model is that $\mathbf{y}'\mathbf{y}$ is not always the total sum of squares for the reduced model. For example, if the full model is

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$

and the hypothesis is $b_1 = b_2 + 4$, then the reduced model would be

$$y_i = b_0 + (b_2 + 4)x_{i1} + b_2 x_{i2} + e_i$$

or

$$y_i - 4x_{i1} = b_0 + b_2(x_{i1} + x_{i2}) + e_i. \tag{127}$$

The total sum of squares for this reduced model is $(\mathbf{y} - 4\mathbf{x}_1)'(\mathbf{y} - 4\mathbf{x}_1)$ and not $\mathbf{y}'\mathbf{y}$ and so $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is not the reduction in the sum of squares. Furthermore, (127) is not the only reduced model because the hypothesis $b_1 = b_2 + 4$ could just as well be used to amend the model to be

$$y_i = b_0 + b_1 x_{i1} + (b_1 - 4)x_{i2} + e_i$$

or

$$y_i + 4x_{i2} = b_0 + b_1(x_{i1} + x_{i2}) + e_i. \tag{128}$$

The total sum of squares will now be $(\mathbf{y} + 4\mathbf{x}_2)'(\mathbf{y} + 4\mathbf{x}_2)$. As a result, in this case, there are two reduced models (127) and (128). They and their total sum of squares are not identical. Neither of the total sum of squares equal $\mathbf{y}'\mathbf{y}$. Therefore, $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is not the reduction in the sum of squares from fitting the reduced model. Despite this, $\text{SSE} + Q$ is the residual sum of squares for all possible reduced models. The total sums of squares and reductions in sums of squares differ from model to model but the residual sums of squares are all the same.

The situation just described is true in general for the hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$. Suppose that $\mathbf{L}'$ is such that $\mathbf{R} = \begin{bmatrix} \mathbf{K}' \\ \mathbf{L}' \end{bmatrix}$ has full rank and $\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{P} & \mathbf{S} \end{bmatrix}$ is its inverse. Then the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ can be written

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{R}^{-1}\mathbf{R}\mathbf{b} + \mathbf{e} \\ &= \mathbf{X}\begin{bmatrix} \mathbf{P} & \mathbf{S} \end{bmatrix}\begin{bmatrix} \mathbf{K}'\mathbf{b} \\ \mathbf{L}'\mathbf{b} \end{bmatrix} + \mathbf{e} \\ &= \mathbf{X}\mathbf{P}\mathbf{m} + \mathbf{X}\mathbf{S}\mathbf{L}'\mathbf{b} + \mathbf{e} \end{aligned}$$

so that

$$\mathbf{y} - \mathbf{X}\mathbf{P}\mathbf{m} = \mathbf{X}\mathbf{S}\mathbf{L}'\mathbf{b} + \mathbf{e}. \tag{129}$$

This is a model in the elements of $\mathbf{L}'\mathbf{b}$ which represents $r - s$ LIN functions of the elements of $\mathbf{b}$. However, since $\mathbf{L}'$ is arbitrarily chosen to make $\mathbf{R}$ non-singular, the model (129) is not unique. In spite of this, it can be shown that the residual sum of squares after fitting any one of the models implicit in (129) is SSE + $Q$. The corresponding value of the estimator of $\mathbf{b}$ is $\hat{\mathbf{b}}_c$ given in (118).

**(*ii*) *The Hypothesis* $\mathbf{K}'\mathbf{b} = \mathbf{0}$** One case where $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ *is* a reduction in the sum of squares due to fitting the reduced model is when $\mathbf{m} = \mathbf{0}$. In this instance, (129) becomes

$$\mathbf{y} = \mathbf{X}\mathbf{S}\mathbf{L}'\mathbf{b} + \mathbf{e}.$$

The total sum of squares for the reduced model is then $\mathbf{y}'\mathbf{y}$, the same as that for the full model. Hence in this case,

$$\mathbf{y}'\mathbf{y} - (\text{SSE} + Q) = \text{reduction(reduced)}. \tag{130}$$

We show that (130) is positive semi-definite and, as a result, a sum of squares. To do this, substitute $\mathbf{m} = \mathbf{0}$ into (126). We have that

$$\begin{aligned} &\mathbf{y}'\mathbf{y} - (\text{SSE} + Q) \\ &\quad = \hat{\mathbf{b}}\mathbf{X}'\mathbf{y} - \hat{\mathbf{b}}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'\hat{\mathbf{b}} \\ &\quad = \mathbf{y}'\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right\}\mathbf{y}. \end{aligned} \tag{131}$$

The matrix of the quadratic form in (131) in curly brackets is idempotent and is thus positive semi-definite so that $\mathbf{y}'\mathbf{y} - (\text{SSE} + Q)$ is a sum of squares. From (130)

$$Q = \mathbf{y}'\mathbf{y} - \text{SSE-reduction(reduced)}.$$

However,

$$\mathbf{y}'\mathbf{y} - \text{SSE} = \text{SSR} = \text{reduction(full)}$$

and so

$$Q = \text{reduction (full)} - \text{reduction(reduced)}.$$

**TABLE 3.6 Analysis of Variance for Testing the Hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Regression(full model) | $r$ | SSR |
| Hypothesis | $s$ | $Q$ |
| Reduced model | $r - s$ | SSR$-Q$ |
| Residual Error | $N - r$ | SSE |
| Total | $N$ | SST |

Therefore, since the sole difference between the full and reduced models is just the hypothesis, it is logical to describe $Q$ as the reduction in the sum of squares due to the hypothesis. With this description, we insert the partitioning of SSR as the sum of $Q$ and SSR – $Q$ into the analysis of variance of Table 3.2 to obtain Table 3.6. In doing so, we utilize (114), that when $\mathbf{m} = \mathbf{0}$,

$$\frac{Q}{\sigma^2} \sim \chi^{2'}\left\{s, \frac{\mathbf{b}'\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'\mathbf{b}}{2\sigma^2}\right\}.$$

Then, because

$$\frac{\mathbf{y}'\mathbf{y} - \text{SSE}}{\sigma^2} \sim \chi^{2'}\left\{r, \frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{2\sigma^2}\right\},$$

an application of Theorem 8, Chapter 2, shows that

$$\frac{\text{SSR} - Q}{\sigma^2} \sim \chi^{2'}\left\{r - s, \frac{\mathbf{b}'[\mathbf{X}'\mathbf{X} - \mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}']\mathbf{b}}{2\sigma^2}\right\}$$

and is independent of $\text{SSE}/\sigma^2$. This, of course, can be derived directly from (131). Furthermore, the non-centrality parameter in the distribution of SSR – $Q$ can in terms of (129) be shown to be equal to $\mathbf{b}'\mathbf{L}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})\mathbf{L}'\mathbf{b}/2\sigma^2$ (see Exercise 15). Hence, under the null hypothesis, this non-centrality parameter is zero when $\mathbf{L}'\mathbf{b} = \mathbf{0}$. Thus, SSR – $Q$ forms the basis of an $F$-test for the sub-hypothesis $\mathbf{L}'\mathbf{b} = \mathbf{0}$ under the null hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$.

We now have the following $F$-tests:

1. To test the full model, we have

$$F = \frac{\text{SSR}/r}{\text{SSE}/(N - r)};$$

2. A test of the hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$ is

$$F = \frac{Q/s}{\text{SSE}/(N - r)};$$

**TABLE 3.7 Analysis Of Variance for Testing the Hypothesis $\mathbf{b}_q = \mathbf{0}$**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Full model | $r$ | SSR $= \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$ |
| Hypothesis $\mathbf{b}_q = \mathbf{0}$ | $q$ | $Q = \hat{\mathbf{b}}_q'\mathbf{T}_{qq}^{-1}\mathbf{b}_q$ |
| Reduced model | $r - q$ | SSR $- Q$ |
| Residual error | $N - r$ | SSE $=$ SST $-$ SSR |
| Total | $N$ | SST $= \mathbf{y}'\mathbf{y}$ |

3. For the reduced model $\mathbf{y} = \mathbf{XSL}'\mathbf{b} + \mathbf{e}$

$$F = \frac{(\text{SSR} - Q)/(r - s)}{\text{SSE}/(N - r)}$$

tests the sub-hypothesis $\mathbf{L}'\mathbf{b} = 0$.

**(*iii*) *The Hypothesis $b_q = 0$*** This is the most useful case of the reduced model when $\mathbf{m} = \mathbf{0}$. For this case, we have $\mathbf{K}' = \begin{bmatrix}\mathbf{I}_q & \mathbf{0}\end{bmatrix}$ for some $q \leq k$. The null hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ then reduces to $\mathbf{b}_q = \mathbf{0}$ where $\mathbf{b}_q' = \begin{bmatrix} b_0 & b_1 & \cdots & b_{q-1}\end{bmatrix}$, say, a subset of $q$ of the $b$'s. We discussed this situation earlier in Section 6c. We found in (122) that

$$F(H) = \frac{Q}{q\hat{\boldsymbol{\sigma}}^2} \text{ with } Q = \mathbf{b}_q'\mathbf{T}_{qq}^{-1}\mathbf{b}_q,$$

involving the "invert part of the inverse" rule. Hence, a special case of Table 3.6 is the analysis of variance for testing the hypothesis $H : \mathbf{b}_q = \mathbf{0}$ shown in Table 3.7.

Table 3.7 shows the most direct way of computing its parts. They are SSR $= \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$, $Q = \hat{\mathbf{b}}_q'\mathbf{T}_{qq}^{-1}\mathbf{b}_q$, SSR $- Q$ by differencing, SST $= \mathbf{y}'\mathbf{y}$, and SSE by differencing. Although SSR $- Q$ is obtained most readily by differencing, it can also be expressed as $\hat{\mathbf{b}}_{cp}'\mathbf{x}_p'\mathbf{X}_p\mathbf{b}_{cp}$ (see Exercise 16). The estimator $\hat{\mathbf{b}}_{cp}$ is derived from (118) as

$$\hat{\mathbf{b}}_{cp} = \hat{\mathbf{b}}_p - \mathbf{T}_{pq}\mathbf{T}_{qq}^{-1}\hat{\mathbf{b}}_q \tag{132}$$

using $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K} = \mathbf{T}_{qq}$ as in (122).

**Example 17 Analysis of Variance for a Test of the Hypothesis $\mathbf{b_q = 0}$** Consider the data for Example 3. We shall test the hypothesis $H$: $b_0 = 0$, $b_1 = 0$ and make an analysis of variance table like Table 3.7 using the results in Table 3.2. From Table 3.2, we have, SSR $= 1042.5$, SSE $= 11.5$, and SST $= 1054$. We have,

$$Q = \begin{bmatrix}2.3333 & 2.0833\end{bmatrix}\begin{bmatrix}17.5889 & 0.6389 \\ 0.6389 & 0.1389\end{bmatrix}\begin{bmatrix}2.3333 \\ 2.0833\end{bmatrix} = 102.5732$$

and SSR – $Q$ = 1042.5 –102.5732 = 939.9268. The analysis of variance table is below.

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Full model | 3 | 1042.5 | 347.5 | 60.43 |
| Hypothesis $b_0 = 0, b_1 = 0$ | 2 | 102.57 | 51.285 | 8.92 |
| Reduced model ($b_2$) | 1 | 939.93 | 939.93 | 163.466 |
| Residual error | 2 | 11.5 | 5.75 | |
| Total | 5 | 1054 | | |

For the full model, we reject $H$:$\mathbf{b} = \mathbf{0}$ at $\alpha = .05$ but not at $\alpha = .01$. The $p$-value is 0.016.

We fail to reject the hypothesis $b_0 = 0$, $b_1 = 0$ at $\alpha = .05$. The $p$-value is 0.10. For the reduced model $y_{i=} b_2 x_{2i}$, we reject the hypothesis $b_2 = 0$ at $\alpha = .05$ and at $\alpha = .01$, the $p$-value being 0.006. □

**Example 18 Illustrations of Tests of Hypothesis for $H$:$\mathbf{K'b} = \mathbf{m}$** Consider the following data

| $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 8 | 2 | 1 | 4 |
| 10 | −1 | 2 | 1 |
| 9 | 1 | −3 | 4 |
| 6 | 2 | 1 | 2 |
| 12 | 1 | 4 | 6 |

We consider no-intercept models. The $\mathbf{X}$ matrix thus consists of the last three columns of the above table. The $\mathbf{y}$ vector is the first column. We have that

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 11 & 3 & 21 \\ 3 & 31 & 20 \\ 21 & 20 & 73 \end{bmatrix} = \begin{bmatrix} 0.2145 & 0.0231 & -0.0680 \\ 0.0231 & 0.0417 & -0.0181 \\ -0.0680 & -0.0181 & 0.0382 \end{bmatrix}.$$

Furthermore,

$$\mathbf{y'y} = 425 \quad \text{and} \quad \mathbf{X'y} = \begin{bmatrix} 39 \\ 55 \\ 162 \end{bmatrix}.$$

Then

$$\mathbf{b'} = \begin{bmatrix} -1.3852 & 0.2666 & 2.5446 \end{bmatrix}.$$

The analysis of variance is

| Source | Degrees of Freedom | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Full model | 3 | 372.9 | 124.3 | 4.77 |
| Residual error | 2 | 52.1 | 26.05 | |
| Total | 5 | 425.0 | | |

We would fail to reject the hypothesis $\mathbf{b} = \mathbf{0}$ at $\alpha = .05$. The $p$-value of the $F$-statistic is 0.177.

We now test the hypothesis $H : b_1 - b_2 = 4$ using (114). We have $\mathbf{K}' = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$ so

$$\mathbf{K}'\hat{\mathbf{b}} - 4 = -.6517 \text{ and} (\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1} = 4.7641$$

and, as a result, $Q = 152.174$. The $F$-statistic is $152.174/(52.1/2) = 5.84$. The $p$-value is 0.136, so we fail to reject the null hypothesis at $\alpha = .05$.

A reduced model where $b_1$ is replaced by $b_2 + 4$ would be

$$y - 4x_1 = b_2(x_1 + x_2) + b_3x_3 + e \tag{133}$$

The data for this model would be

| $y - 4x_1$ | $x_1 + x_2$ | $x_3$ |
|---|---|---|
| 0 | 3 | 4 |
| 14 | 1 | 1 |
| 5 | –2 | 4 |
| –2 | 3 | 2 |
| 8 | 5 | 6 |

The total sum of squares $(y - 4x_1)'(y - 4x_1) = 289$. The residual sum of squares, using SSE from the analysis of variance and $Q$ from the $F$-statistic is

| Source | d.f. | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Regression (reduced model) | 2 | 84.7 | 42.35 | 0.622 |
| Residual error | 3 | 204.3 | 68.1 | |
| Total | 5 | 289.0 | | |

The $F$-statistic being less than unity is not significant.

The value of 84.7 for the reduction sum of squares may also be obtained by using the normal equations for (133). In matrix form the normal equations are

$$\begin{bmatrix} 48 & 41 \\ 41 & 73 \end{bmatrix} \begin{bmatrix} \hat{b}_2^c \\ \hat{b}_3^c \end{bmatrix} = \begin{bmatrix} 38 \\ 78 \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} \hat{b}_2^c \\ \hat{b}_3^c \end{bmatrix} = \frac{1}{1823} \begin{bmatrix} 73 & -41 \\ -41 & 48 \end{bmatrix} \begin{bmatrix} 38 \\ 78 \end{bmatrix} = \begin{bmatrix} -0.2326 \\ 1.1991 \end{bmatrix}.$$

Then the reduction sum of squares is

$$\begin{bmatrix} -0.2326 & 1.1991 \end{bmatrix} \begin{bmatrix} 38 \\ 78 \end{bmatrix} = 84.691,$$

which is the same as the 84.7 in the analysis of variance table rounding to the nearest tenth.

These calculations are shown here purely to illustrate the sum of squares in the analysis of variance. They are not needed specifically because, for the reduced model, the residual is always SSE + Q. The estimator of **b** may be found from (118) as

$$\begin{aligned} \hat{\mathbf{b}}_c = & \begin{bmatrix} -1.3852 \\ 0.2666 \\ 2.5446 \end{bmatrix} - \begin{bmatrix} 0.2145 & 0.0231 & -0.0680 \\ 0.0231 & 0.0417 & -0.0181 \\ -0.0680 & -0.0181 & 0.0382 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \cdot \\ & \left[ \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0.2145 & 0.0231 & -0.0680 \\ 0.0231 & 0.0417 & -0.0181 \\ -0.0680 & -0.0181 & 0.0382 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right]^{-1} \cdot \\ & \left( \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} -1.3852 \\ 0.2666 \\ 2.5446 \end{bmatrix} - [4] \right) \\ = & \begin{bmatrix} 3.7674 \\ -0.2326 \\ 1.1991 \end{bmatrix}, \end{aligned}$$

where $\hat{b}_1^c - \hat{b}_2^c = 4$ and $\hat{b}_2^c$ and $\hat{b}_3^c$ are as before.

For testing the hypothesis $b_1 = 0$, we have that $\mathbf{K}' = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ so $[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1} = \frac{1}{0.2145}$ and $\mathrm{Q} = \frac{(-1.3852)^2}{0.2145} = 8.945$. The analysis of variance of Table 3.6 is

| Source | d.f. | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Full model | 3 | 372.9 | 124.3 | 4.77 |
| Hypothesis | 1 | 8.9 | 8.9 | 0.342 |
| Reduced model | 2 | 364.0 | 182 | 6.99 |
| Residual error | 2 | 52.1 | 26.05 | |
| Total | 5 | 425 | | |

None of the effects are statistically significant. The $p$-values for the $F$-statistic are for the full model 0.178, the hypothesis 0.617, and the reduced model 0.125. The restricted least-square estimator is

$$\hat{\mathbf{b}}_c = \begin{bmatrix} 1.3852 \\ 0.2666 \\ 2.5466 \end{bmatrix} - \begin{bmatrix} 0.2145 \\ 0.0231 \\ -0.0680 \end{bmatrix} \frac{(-1.3852)}{0.2145} = \begin{bmatrix} 0 \\ 0.4158 \\ 2.1074 \end{bmatrix}.$$

These results may be verified using the normal equations of the reduced model. In this case, we have

$$\begin{bmatrix} 31 & 20 \\ 20 & 73 \end{bmatrix} \begin{bmatrix} \hat{b}_{2c} \\ \hat{b}_{3c} \end{bmatrix} = \begin{bmatrix} 55 \\ 162 \end{bmatrix}.$$

As a result,

$$\begin{bmatrix} \hat{b}_{2c} \\ \hat{b}_{3c} \end{bmatrix} = \frac{1}{1863} \begin{bmatrix} 73 & -20 \\ -20 & 31 \end{bmatrix} \begin{bmatrix} 55 \\ 162 \end{bmatrix} = \begin{bmatrix} 0.4160 \\ 2.1052 \end{bmatrix}.$$

the same result with a slight error due to rounding off. Also,

$$\text{SS(ReducedModel)} = \begin{bmatrix} 0.4160 & 2.1052 \end{bmatrix} \begin{bmatrix} 31 & 20 \\ 20 & 73 \end{bmatrix} \begin{bmatrix} 0.4160 \\ 2.1052 \end{bmatrix} = 363.9,$$

rounded to 364. □

We now consider another example where we will use SAS software to fit a regression model and test a hypothesis about the regression coefficients. These data will also be used to illustrate multicollinearity in Section 6e.

**Example 19 Growth Rates in Real Gross Domestic Product for the United States and Several European Countries, 2004–2013**

| Country | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|---|---|---|---|---|
| United States | 3.5 | 3.1 | 2.7 | 1.9 | −0.3 | −3.1 | 2.4 | 1.8 | 2.3 | 2.0 |
| Germany | 0.7 | 0.8 | 3.9 | 3.4 | 0.8 | −5.1 | 4.0 | 3.1 | 0.9 | 0.6 |
| France | 2.5 | 1.8 | 2.5 | 2.3 | −0.1 | −3.1 | 1.7 | 1.7 | 0.2 | 0.3 |
| Italy | 1.7 | 0.9 | 2.2 | 1.7 | −1.2 | −5.5 | 1.8 | 0.4 | −2.1 | −1.0 |
| Spain | 3.3 | 3.6 | 4.1 | 3.5 | 0.9 | −3.7 | −0.3 | 0.4 | −1.4 | −1.5 |

Using SAS, we fit a regression line with response variable the United States and predictor variables the four European countries and tested the hypothesis $H : b_2 + 2(b_1 + b_3 + b_4) = 0$. The output follows:

**The SAS System**

**The GLM Procedure**

| | |
|---|---|
| Number of Observations Read | 10 |
| Number of Observations Used | 10 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable United**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 30.67825467 | 7.66956367 | 10.95 | 0.0109 |
| Error | 5 | 3.50274533 | 0.70054907 | | |
| Corrected Total | 9 | 34.18100000 | | | |

| R − Square | Coeff Var | Root MSE | United Mean |
|---|---|---|---|
| 0.897524 | 51.34896 | 0.836988 | 1.630000 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Germany | 1 | 18.64234395 | 18.64234395 | 26.61 | 0.0036 |
| France | 1 | 9.28469640 | 9.28469640 | 13.25 | 0.0149 |
| Italy | 1 | 1.087872276 | 1.087872276 | 1.54 | 0.2697 |
| Spain | 1 | 1.67249157 | 1.67249157 | 2.39 | 0.1830 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Germany | 1 | 0.66619311 | 0.66619311 | 0.95 | 0.3743 |
| France | 1 | 6.34162434 | 6.34162434 | 9.05 | 0.0298 |
| Italy | 1 | 0.42624103 | 0.42624103 | 0.61 | 0.4707 |
| Spain | 1 | 1.67249157 | 1.67249157 | 2.39 | 0.1830 |

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| france+2(germany + italy + spain) = 0 | 1 | 0.33229619 | 0.33229619 | 0.47 | 0.5216 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | −0.184349768 | 0.90824799 | −0.20 | 0.8472 |
| Germany | −0.253669749 | 0.26012847 | −0.98 | 0.3743 |
| France | 2.463058021 | 0.81864133 | 3.01 | 0.0298 |
| Italy | −0.473991811 | 0.60766257 | −0.78 | 0.4707 |
| Spain | −0.358740247 | 0.23217605 | −1.55 | 0.1830 |

The code to generate this output was

```
data growth;
input United Germany France Italy Spain;
datalines;
proc glm;
model United=Germany France Italy Spain;
contrast 'france +2(Germany +Italy +spain)=0'france 1
germany 2 italy 2 spain 2;
run;
```

The best fitting model was

$$y = -0.1843 - 0.2537x_1 + 2.4631x_2 - 0.4740x_3 - 0.3587x_4.$$

It accounts for 89% of the variation. The type I sum of squares is the sum of squares when the variables are added sequentially. For example, the type I sum of squares for France is the difference between the sum of squares for the model with Germany and France and the sum of squares for France alone. All of these type I sums of squares add up to the model sum of squares. Had the variables been fitted in a different order, the type I sums of squares would have been different but would have still added up to the model sum of squares. The type III sum of squares is the sum of squares of the variables given that of the other variables. It is as if that variable was added to the model last. The type III sums of squares will be the same regardless of the order that the variables were added. For this example, the type I sum of squares is statistically significant for Germany and France but not the other variables, given the order in which they were added Germany, France, Spain, Italy. The only significant sum for type III is France. There are 24(4!) different orders in which the variables could have been added. The reader might like to compare the results for some of them. □

### e. Stochastic Constraints

In Section 6d, we considered exact constraints of the form $\mathbf{K}'\mathbf{b} = \mathbf{m}$. We will now let $\mathbf{K}' = \mathbf{R}$ and $\mathbf{m} = \mathbf{r}$ and consider stochastic constraints of the form $\mathbf{r} = \mathbf{R}\mathbf{b} + \eta$, where the elements of the vector $\eta$ are independent with mean zero and variance $\tau^2$. We now consider the augmented model

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{R} \end{bmatrix} \mathbf{b} + \begin{bmatrix} \mathbf{e} \\ \tau \end{bmatrix}, \tag{134}$$

where, as before, the elements of $\mathbf{e}$ are independent with mean zero and variance $\sigma^2$. Finding the weighted least-square estimator by minimizing,

$$m = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})}{\boldsymbol{\sigma}^2} + \frac{(\mathbf{r} - \mathbf{R}\mathbf{b})'(\mathbf{r} - \mathbf{R}\mathbf{b})}{\boldsymbol{\tau}^2}.$$

Differentiation with respect to $\mathbf{b}$ yields the normal equations in matrix form

$$(\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})\hat{\mathbf{b}}_m = \tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}$$

with solution the mixed estimator of Theil and Goldberger (1961),

$$\hat{\mathbf{b}}_m = (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^{-1}(\tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}). \tag{135}$$

The constraint $\mathbf{r} = \mathbf{R}\mathbf{b} + \eta$ may be thought of as stochastic prior information or as taking additional observations. Notice that

$$\begin{aligned}\hat{\mathbf{b}}_m &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^{-1}(\tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}) \\ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^{-1}(\tau^2\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{r}) \\ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^{-1}(\tau^2\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}_1 + \sigma^2\mathbf{R}'\mathbf{R}\hat{\mathbf{b}}_2),\end{aligned}$$

where $\hat{\mathbf{b}}_1 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\mathbf{b}}_2 = (\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{r}$, the least-square estimators for each of the two models in the augmented model (134).

**Example 20 A Mixed Estimate** We will fit regression models for the data of Example 19 using only the variables for France and Germany and obtain a mixed estimator. We use the first 5 years as prior information and the last 5 years as sample information. Thus, we have

$$\mathbf{y}' = \begin{bmatrix}-3.1 & 2.4 & 1.8 & 2.3 & 2.0\end{bmatrix} \quad \text{and} \quad \mathbf{r}' = \begin{bmatrix}3.5 & 3.1 & 2.7 & 1.9 & -0.3\end{bmatrix}.$$

We also have that

$$\mathbf{X} = \begin{bmatrix}1 & -5.1 & -3.1 \\ 1 & 4.0 & 1.7 \\ 1 & 3.1 & 1.7 \\ 1 & 0.9 & 0.2 \\ 1 & 0.6 & 0.3\end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix}1 & 0.7 & 2.5 \\ 1 & 0.8 & 1.8 \\ 1 & 3.9 & 2.5 \\ 1 & 3.4 & -2.3 \\ 1 & 0.8 & -0.1\end{bmatrix}.$$

Furthermore, as a result,

$$\hat{\mathbf{b}}_1' = \begin{bmatrix}0.8999 & 0.0096 & 1.08373\end{bmatrix} \quad \text{and} \quad \hat{\mathbf{b}}_2' = \begin{bmatrix}1.3315 & 0.2302 & 0.4620\end{bmatrix}.$$

Since $\sigma^2$ and $\tau^2$ are unknown, we estimate them by using the formulae

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_1)'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_1)}{N - r} \quad \text{and} \quad \hat{\tau}^2 = \frac{(\mathbf{r} - \mathbf{R}\hat{\mathbf{b}}_2)'(\mathbf{r} - \mathbf{R}\hat{\mathbf{b}}_2)}{N - r}$$

with $N = 5, r = 3$. Then $\hat{\sigma}^2 = 1.0755$ and $\hat{\tau}^2 = 2.8061$. Using these estimates together with (135), we get

$$\hat{\mathbf{b}}_m' = \begin{bmatrix} 0.8931 & 0.3262 & 0.5182 \end{bmatrix}.$$

Computing the least-square estimator using all 10 observations, we have

$$\hat{\mathbf{b}}' = \begin{bmatrix} 0.9344 & 0.3267 & 0.5148 \end{bmatrix}.$$

The reason for the slight difference in the two estimators is the mixed estimator is a weighted least-square estimator using the estimates of the variance as weights.

### f. Exact Quadratic Constraints (Ridge Regression)

In Section 6b, we considered finding the least-square estimator subject to a linear constraint. We now consider the problem of finding the least-square estimator subject to a quadratic constraint. We shall minimize $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ subject to the constraint $\mathbf{b}'\mathbf{H}\mathbf{b} = \phi_o$, where $\mathbf{H}$ is a positive definite matrix. As was done in Section 6b, we employ Lagrange multipliers. To this end, we write

$$L = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + \lambda(\mathbf{b}'\mathbf{H}\mathbf{b} - \phi_0),$$

obtain its derivative and set it equal to zero and solve for $\hat{\mathbf{b}}_r$. Thus, we have

$$\frac{\partial L}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} + 2\lambda\mathbf{H}\mathbf{b} = \mathbf{0}$$

and

$$\hat{\mathbf{b}}_r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}'\mathbf{y}.$$

Let $\mathbf{G} = \lambda\mathbf{H}$ and we obtain the generalized ridge regression estimator of Rao (1975)

$$\hat{\mathbf{b}}_r = (\mathbf{X}'\mathbf{X} + \mathbf{G})^{-1}\mathbf{X}'\mathbf{y}. \tag{136}$$

When $\mathbf{G} = k\mathbf{I}$, the estimator in (136) reduces to the estimator of Hoerl and Kennard (1970)

$$\hat{\mathbf{b}}_r = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

Ridge regression estimators are especially useful for data where there is a linear relationship between the variables. Such data are called multicollinear data. When an exact linear relationship exists between the variables, the $\mathbf{X}'\mathbf{X}$ matrix has less

than full rank and a possible solution is to use the least-square estimator $\mathbf{b} = \mathbf{G}\mathbf{X}'\mathbf{y}$, where $\mathbf{G}$ is a generalized inverse. On the other hand, the $\mathbf{X}'\mathbf{X}$ may have very small eigenvalues. In this instance, the total variance

$$TV = \sigma^2 Tr(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i}.$$

may be very large. We now give an example.

**Example 21 Improving the Precision by Using Ridge Estimators** We shall use the data from Example 18. One reason to suspect that ridge estimators might be useful is the fact that there is a high correlation between the growth rates of France and Italy, 0.978. The correlation matrix is

$$\mathbf{R} = \begin{bmatrix} 1 & 0.868 & 0.880 & 0.607 \\ 0.868 & 1 & 0.978 & 0.846 \\ 0.880 & 0.978 & 1 & 0.836 \\ 0.607 & 0.846 & 0.826 & 1 \end{bmatrix}.$$

Ridge estimators are usually obtained using standardized data. This is obtained for each country by subtracting the mean and dividing by the standard deviation. These are

| United States | Germany | France | Italy | Spain |
|---|---|---|---|---|
| 0.95956 | −0.22845 | 0.87829 | 0.75439 | 0.90411 |
| 0.75430 | −0.19100 | 0.47831 | 0.42096 | 1.01666 |
| 0.54905 | 0.96997 | 0.87829 | 0.96279 | 1.20423 |
| 0.13855 | 0.78272 | 0.76272 | 0.75439 | 0.97914 |
| −0.99034 | −0.19100 | −0.62405 | −0.45430 | 0.00375 |
| −2.42711 | −2.40058 | −2.35751 | −2.24651 | −1.7294 |
| 0.39511 | 1.00742 | 0.41603 | 0.79607 | −0.44643 |
| 0.08723 | 0.67037 | 0.41603 | 0.21256 | −0.18382 |
| 0.34380 | −0.15355 | −0.45070 | −0.82941 | −0.85909 |
| 0.18986 | −0.26590 | −0.39292 | −0.37094 | −0.89661 |

The least-square fit would be

$$y = -0.3476x_1 + 2.1673x_2 - 0.5836x_3 - 0.4907x_4.$$

The total variance of the least-square estimator would be $7.135\sigma^2$.

To give a good ridge regression estimator, we need to estimate the parameter $k$. We would like to get the ridge estimator with the smallest possible mean square error. Many ways of doing this are proposed in the literature. One method is to plot the coefficients of the ridge estimator for different values of the parameter $k$ and see

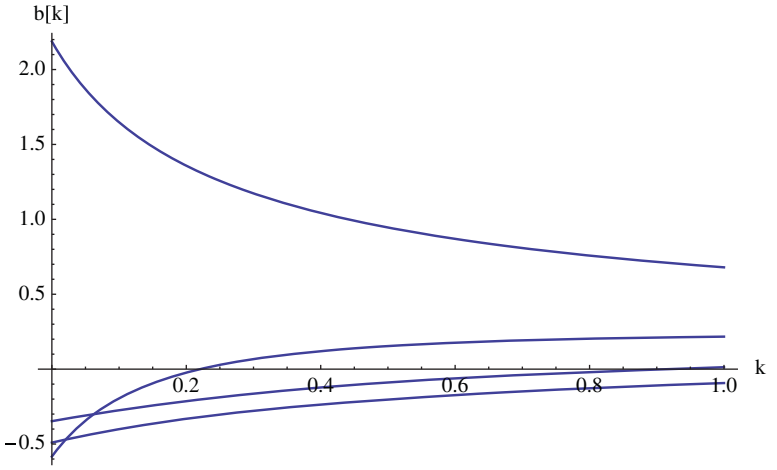



**FIGURE 3.4** The Ridge Trace

where it appears to stabilize. This is called the ridge trace shown in Figure 3.4. It appears to stabilize around $k = 0.15$.

Another method is to use the point estimate $\hat{k} = m\hat{\sigma}^2/\mathbf{b}'\mathbf{b}$ suggested by Hoerl and Kennard (1970). The rationale for this method is to use a generalized ridge estimator with a diagonal matrix K, reparameterize the model by orthogonal transformations to one where the $\mathbf{X}'\mathbf{X}$ matrix is diagonal and show that the mean square error of the individual coefficients is smallest for $\sigma^2/(\text{coefficient})^2$ and find reciprocal of the mean. See Hoerl and Kennard (1970) or Gruber (1998) for more details. In this instance, the estimate of $k$ is 0.112. The fit using the ridge estimator for this estimate of $k$ would be

$$y = -0.2660x_1 + 1.6041x_2 - 0.1666x_3 - 0.3904x_4.$$

The total variance of the ridge estimator for this value of $k$ would be $4.97984\sigma^2$, a substantial reduction.

More information about ridge type estimators is available in Gruber (1998, 2010).

## 7. RELATED TOPICS

It is appropriate to briefly mention certain topics related to the preceding developments that are customarily associated with testing hypothesis. The treatment of these topics will do no more than act as an outline to the reader showing him or her their application to the linear models procedure. As with the discussion of distribution functions, the reader will have to look elsewhere for a complete discussion of these topics. A

comprehensive treatment of hypothesis testing is available, for example, in Lehmann and Romano (2005).

### a. The Likelihood Ratio Test

Tests of linear hypotheses $\mathbf{K}'\mathbf{b} = \mathbf{m}$ have already been developed from the starting point of the $F$-statistic. This, in turn, can be shown to stem from the likelihood ratio test.

For a sample of $N$ observations $\mathbf{y}$, where $\mathbf{y} \sim N(\mathbf{Xb}, \sigma^2\mathbf{I})$, the likelihood function is

$$L(\mathbf{b}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}N} \exp\left\{-\left[\frac{(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{2\sigma^2}\right]\right\}.$$

The likelihood ratio test utilizes two values of $L(\mathbf{b}, \sigma^2)$:

(i) The maximum value of $L(\mathbf{b}, \sigma^2)$ maximized over the complete range of parameters, $0 < \sigma^2 < \infty$ and $-\infty < b_i < \infty$ for all $i$, $\max(L_w)$;
(ii) The maximum value of $L(\mathbf{b}, \sigma^2)$ maximized over the range of parameters limited restricted or defined by the hypothesis $H$,$\max(L_H)$.

The ratio of the two maxima,

$$L = \frac{\max(L_H)}{\max(L_w)},$$

is called the likelihood ratio. Each maximum is found in the usual manner.

(i) Differentiate $L(\mathbf{b}, \sigma^2)$ with respect to $\sigma^2$ and the elements of $\mathbf{b}$.
(ii) Equate the differentials to zero.
(iii) Solve the resulting equations for $\mathbf{b}$ and $\sigma^2$.
(iv) Use these equations in place of $\mathbf{b}$ and $\sigma^2$ in $L(\mathbf{b}, \sigma^2)$.

For the case of $L_{\mathrm{H}}$, carry out the maximization within the limits of the hypothesis.

We demonstrate the procedure outlined above for the case of the hypothesis $H$: $\mathbf{b} = \mathbf{0}$.

First, as we have seen, $\partial L(\mathbf{b}, \sigma^2)/\partial \mathbf{b} = \mathbf{0}$ gives $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\partial L(\mathbf{b}, \sigma^2)/\partial \sigma^2 = 0$ gives $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})/N$. Thus,

$$\begin{aligned}\max(L_w) = L(\hat{\mathbf{b}}, \hat{\sigma}^2) &= (2\pi\hat{\sigma}^2)^{-\frac{1}{2}N} \exp\left\{-\left[\frac{(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})}{2\hat{\sigma}^2}\right]\right\} \\ &= \frac{e^{-\frac{1}{2}N} N^{\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}N}[(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})]^{\frac{1}{2}N}}.\end{aligned}$$

This is the denominator of $L$. The numerator comes by replacing $\mathbf{b}$ by $\mathbf{0}$ in the likelihood function. We obtain

$$L(0, \hat{\sigma}^2) = (2\pi\sigma^2)^{-\frac{1}{2}N} \exp - \left(\frac{\mathbf{y}'\mathbf{y}}{2\sigma^2}\right).$$

Maximize this with respect to $\sigma^2$ by solving the equation $\partial L(\mathbf{0}, \sigma^2)/\partial\sigma^2 = 0$. We obtain $\hat{\sigma}^2 = \mathbf{y}'\mathbf{y}/N$. Thus,

$$\begin{aligned}\max(L_H) = L(\mathbf{0}, \hat{\sigma}^2) &= (2\pi\hat{\sigma}^2)^{-\frac{1}{2}N} \exp - \left(\frac{\mathbf{y}'\mathbf{y}}{2\hat{\sigma}^2}\right) \\ &= \frac{e^{-\frac{1}{2}N} N^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}N} (\mathbf{y}'\mathbf{y})^{\frac{1}{2}N}}.\end{aligned}$$

With these values for the maxima, the likelihood ratio is

$$\begin{aligned}L = \frac{\max(L_H)}{\max(L_w)} &= \left[\frac{(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})}{\mathbf{y}'\mathbf{y}}\right]^{\frac{1}{2}N} = \left[\frac{\text{SSE}}{\text{SSR} + \text{SSE}}\right]^{\frac{1}{2}N} \\ &= \left[\frac{1}{1 + \text{SSR}/\text{SSE}}\right]^{\frac{1}{2}N}.\end{aligned}$$

Observe that $L$ is a single-valued function of SSR/SSE that is monotonic decreasing. Therefore, SSR/SSE may be used as a test statistic in place of $L$. By the same reasoning, we can use (SSR/SSE)[(*N*–*r*)/*r*] which follows the $F$-distribution. Thus, we have established the use of the $F$-statistic as an outcome of the likelihood ratio test. In a like manner, we can establish the basis of $F(H)$.

### b. Type I and Type II Errors

When performing a test of hypothesis, there are two ways we can make a mistake. We can reject a null hypothesis when it is true. This is called the type I error. We can accept a false hypothesis. This is called the type II error. The probability of committing a type I error is called the $\alpha$-risk. The probability of committing a type II error is called the $\beta$-risk. We consider these risks in the context of testing the null hypothesis $H$:$\mathbf{K}'\mathbf{b} = \mathbf{m}$.

Under the null hypothesis, $H$:$\mathbf{K}'\mathbf{b} = \mathbf{m}$, $F(R) = (N - r)Q/s\text{SSE}$ has the $F_{s,N-r}$-distribution. If $u$ is any variable having the $F_{s,N-r}$-distribution, then $F_{\alpha,s,N-r}$ is the value where $\Pr\{u \geq F_{\alpha,s,N-r}\} = \alpha$. For a significance test at the 100 $\alpha$ % level the rule of the test is to not reject $H$ whenever $F \leq F_{\alpha,s,N-r}$ and to reject $H$ whenever $F > F_{\alpha,s,N-r}$.

The probability $\alpha$ is the *significance level* of the significance test. As has already been pointed out, popular values of $\alpha$ are 0.05, 0.01, and 0.10. However, there is

nothing sacrosanct about these values. Any value of $\alpha$ between zero and one can be used. The probability of a type I error is the significance level of the test, frequently specified in advance. When we perform a test at $\alpha = .05$, say, we are willing to take a chance of one in 20 of falsely rejecting a true null hypothesis.

Consider the situation where $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ is false but instead some other hypothesis $H_a : \mathbf{K}_a'\mathbf{b}_a = \mathbf{m}$ is true. As in (115)

$$F(H) \sim F'(s, N - r, \lambda) \tag{137}$$

with non-centrality parameter

$$\begin{aligned} \lambda &= \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{2\sigma^2} \\ &= \frac{1}{2}(\mathbf{K}'\mathbf{b} - \mathbf{m})'[\text{var}(\mathbf{K}'\hat{\mathbf{b}})]^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m}) \end{aligned} \tag{138}$$

using (62) for var($\hat{\mathbf{b}}$). Observe that $\lambda \neq 0$ because $\mathbf{K}'\mathbf{b} - \mathbf{m}$ but $\mathbf{K}_a'\mathbf{b} = \mathbf{m}_a$. Suppose that, without our knowing it, the alternative hypothesis $H_a$ was true at the time the data were collected. Suppose that with these data, the hypothesis $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ is tested using $F(H)$ as already described. When $F(H) \leq F_{\alpha,s,N-r}$ we fail to reject $H$. By doing this, we make an error. The error is that we fail to reject $H$ not knowing that $H_a$ was true and $H$ was not true. We fail to reject $H$ when it was false and thus commit a type II error. The $\beta$-risk denoted by $\beta(\lambda)$ for different values of the parameter $\lambda$ is

$$\begin{aligned} \beta(\lambda) = \text{P(II)} &= \Pr\{\text{Type II error occuring}\} \\ &= \Pr\{\text{not rejecting } H \text{ when } H \text{ is false}\} \\ &= \Pr\{F(H) \leq F_{\alpha,s,N-r} \text{ where } F(H) \sim F'(s, N - r, \lambda)\}. \end{aligned} \tag{139}$$

From (136) and (137), we write (139) as

$$\beta(\lambda) = \text{P(II)} = \Pr\{F'(s, N - r, \lambda) \leq F_{\alpha,s.N-r}\}. \tag{140}$$

Equation (140) gives the probability that a random variable distributed as $F'(s, N - r, \lambda)$ is less than $F_{\alpha,s.N-r}$, the $100\alpha\%$ point in the central $F_{s,N\text{-}r}$-distribution. The two kinds of errors are summarized in Table 3.8 below.

As we have already seen to obtain the probabilities of the type II error, we need to obtain values of the non-central distribution. Values are tabulated to help obtain these probabilities in Tang (1938), Kempthorne (1952), and Graybill (1961). For an illustrative example using these tables, see the first edition Searle (1971). We will do a similar example using Mathematica to obtain the desired probabilities.

### c. The Power of a Test

From the expression for the non-centrality parameter $\lambda$, it can be seen that the beta risk $\beta(\lambda)$ of (140) depends upon $\mathbf{K}_a'$ and $\mathbf{m}_a$ of the alternative hypothesis $H_a : \mathbf{K}_a'\mathbf{b} = \mathbf{m}_a$.

**TABLE 3.8 Type I and Type II Errors in Hypothesis Testing**

| | Result of Test of Hypothesis | |
|---|---|---|
| Null hypothesis | $F(H) \le F_{\alpha,s,N-r}$ | $F(H) > F_{\alpha,s,N-r}$ |
| $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ | Conclusion | |
| | Do not reject $H$ | Reject $H$ |
| True | No error | Type I error[a] |
| False($H : \mathbf{K}_a\mathbf{b} = \mathbf{m}_a$ is true) | Type II error[b] | No error |

The probabilities of type I and type II errors are, respectively,
[a]Pr{type I error} $= \alpha = \Pr\{F(H) > F_{\alpha,s,N-r}\}$ when $H : \mathbf{K}'\mathbf{b} = \mathbf{m}$ is true}; $\gamma(\lambda) = 1 - \beta(\lambda)$.
[b]Pr{type II error} $= \beta = \Pr\{F(H) \le F_{\alpha,s,N-r}\}$when $H_a$:$\mathbf{K}_a'\mathbf{b} = \mathbf{m}_a$is true}.

The probability $\gamma(\lambda) = 1 - \beta(\lambda)$ is similarly dependent. It is known as the *power of the test* with respect to the alternative hypothesis. From (139) it is seen that

$$\begin{aligned} \text{Power} &= 1 - \Pr(\text{type II error}) \\ &= 1 - \Pr\{\text{not rejecting } H \text{ when } H \text{ is false}\} \\ &= \Pr\{\text{rejecting } H \text{ when } H \text{ is false}\}. \end{aligned} \tag{141}$$

In other words, the power of the test is the probability you do what you are supposed to when a given alternative hypothesis is the true one; reject the null hypothesis! A test is better than another test if for all values of the parameter $\lambda$ the power is larger. For more information about the power of a test, see Lehmann and Romano (2005).

### d. Estimating Residuals

Residuals are used to determine whether the assumptions that the error terms are independent, have a constant variance, and follow a normal distribution are true. The vector of residuals is the estimated error vector

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}. \tag{142}$$

Some elementary but important properties of residuals are worth mentioning. Recall from (80) that $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The matrix $\mathbf{P}$ is symmetric and idempotent. Furthermore,

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \mathbf{P}\mathbf{y}.$$

We also have that $\mathbf{PX} = \mathbf{0}$. An important and useful property of residuals is that they sum to zero. Recall from (94) that $\mathbf{1}'\mathbf{P} = \mathbf{0}'$. Another important fact is that their sum of squares is SSE as mentioned in (81). Notice that

$$\sum_{i=1}^{N} \hat{e}_i = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{P}'\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \text{SSE}.$$

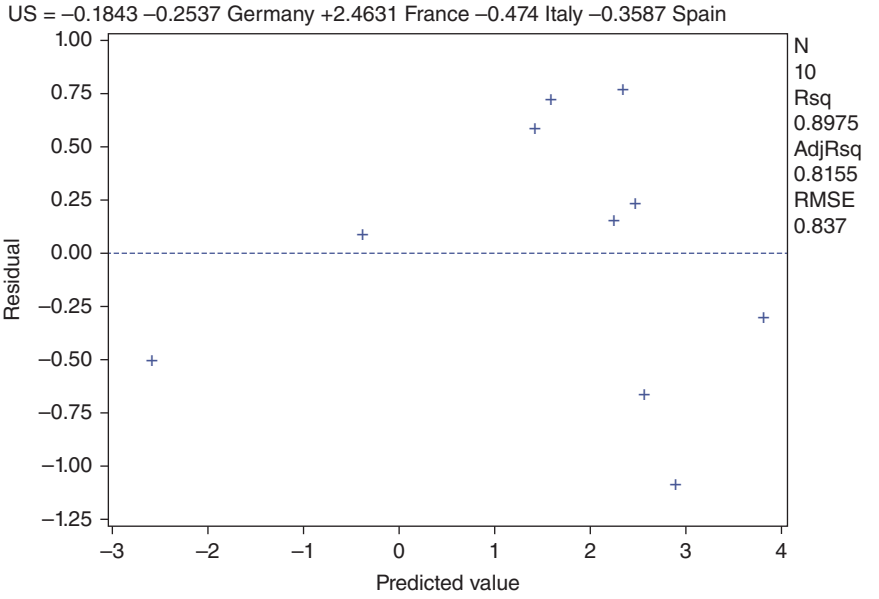



**FIGURE 3.5** Plot of Residuals vs. Predicted Values for Growth Data

Residuals have expected value zero and variance covariance matrix $\mathbf{P}\sigma^2$. Indeed

$$E(\hat{\mathbf{e}}) = E(\mathbf{Py}) = \mathbf{PXb} = \mathbf{0}$$

and

$$\text{var}(\hat{\mathbf{e}}) = \text{var}(\mathbf{Py}) = \mathbf{P}^2\sigma^2 = \mathbf{P}\sigma^2.$$

Additional results will be obtained in the exercises.

The properties just described hold true for the residuals of any intercept model. For example, in assuming normality of the error terms in the model, we have that $\hat{\mathbf{e}} \sim N(\mathbf{0}, \mathbf{P}\sigma^2)$. To determine if there is reason that the normality assumption is not satisfied, one can make a normal probability plot of the $\hat{e}_i$. See for example Figure 3.5. If the values lie far from a straight line, there may be reason to doubt the normality assumption. In doing this, we ignore the fact that $\text{var}(\hat{\mathbf{e}}) = \mathbf{P}\sigma^2$ which means the $\hat{e}_i$ are correlated. Anscombe and Tukey (1963) indicate, for at least a two-way table with more than three rows and columns, "the effect of correlation in graphical procedures is usually negligible." Draper and Smith (1998) provide further discussion of this point.

Other procedures for residual analysis include plotting the residuals against the predicted values of the dependent variables and against the $x$'s. See for example Figure 3.6. Such plots might indicate that the variances of the error terms are not constants or that additional terms, not necessarily linear, are needed in the model. See Draper and Smith (1998) and the references therein for more information.

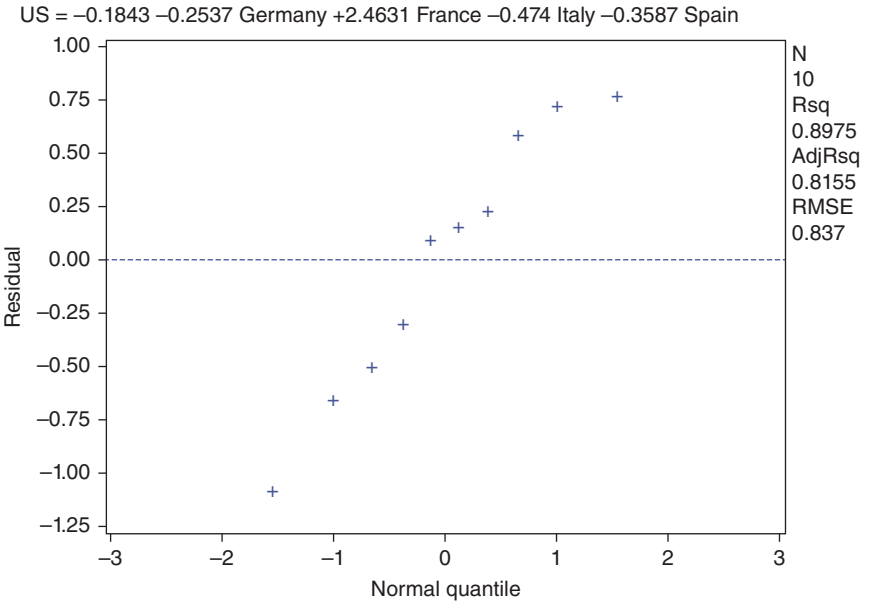



**FIGURE 3.6** Normal Probability Plot for Residuals vs.Growth Data

We now give an illustration of some residual plots.

## 8. SUMMARY OF REGRESSION CALCULATIONS

The more frequently used general expressions developed in this chapter for estimating the linear regression on $k$ $x$ variables are summarized and listed below.

$N$: number of observations on each variable.

$k$: number if $x$ variables.

$y$: $N \times 1$ vector of observed $y$ values.

$\mathbf{X}_1$: $N \times k$ vector of observed $y$ values.

$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_1 \end{bmatrix}$.

$\bar{y}$: mean of the observed $y$ values.

$\bar{\mathbf{x}}' = (1/N)\mathbf{1}'\mathbf{X}_1$: vector of means of observed $x$'s.

$\mathbf{b} = \begin{bmatrix} b_0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}$ : $b_0$ is the intercept; : $\tilde{\boldsymbol{\beta}}$ is vector of regression coefficients.

$\mathcal{X} = \mathbf{X}_1 - \mathbf{1}\bar{\mathbf{x}}'$: matrix of observed $x$'s expressed as deviations from their means.

$\mathcal{X}'\mathcal{X}$: matrix of corrected sums of squares and products of observed $x$'s.

$\mathcal{X}'\mathbf{y}$ : vector of corrected sums of products of observed $x$'s and $y$'s.

$r = k + 1$: rank of $\mathbf{X}$.
$\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$**:** total sum of squares corrected for the mean.
$\hat{\boldsymbol{\theta}} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y}$: estimated regression coefficients.
$\text{SSE} = \text{SST}_\text{m} - \hat{\boldsymbol{\theta}}\mathcal{X}'\mathbf{y}$: error sum of squares.
$\hat{\sigma}^2 = \text{SSE}/(N - r)$: estimated residual error variance.
$\hat{\text{var}}(\hat{\boldsymbol{\theta}}) = (\mathcal{X}'\mathcal{X})^{-1}\hat{\sigma}^2$: estimated covariance matrix of $\hat{\mathbf{b}}$.
$\text{SSR}_\text{m} = \hat{\boldsymbol{\theta}}'\mathcal{X}'\mathbf{y}$: sum of squares due to fitting the model over and above the mean.
$R^2 = \text{SSR/SST}$: coefficient of determination.
$F_{r-1,N-r} = \text{SSR}_\text{m}/(r-1)\hat{\sigma}^2$: $F$-statistic for testing $H : \tilde{\boldsymbol{\theta}} = \mathbf{0}$.
$a^{ii}$ = ith diagonal element of $(\mathcal{X}'\mathcal{X})^{-1}$.
$t_i = \tilde{\boldsymbol{\theta}}_i/\hat{\sigma}\sqrt{a^{ii}}$: $t$-statistic on $N - r$ degrees of freedom for testing hypothesis $H : \tilde{\boldsymbol{\theta}}_i = 0$.
$\hat{\boldsymbol{\theta}}_i \pm t_{N-r,\frac{1}{2}\alpha}\sqrt{a^{ii}\hat{\sigma}^2}$: symmetric $100(1 - \alpha)\%$ confidence interval for $\tilde{b}_i$.
$F_{q,N-r} = \hat{\mathbf{b}}_q'\tilde{\mathcal{T}}_{qq}^{-1}\hat{\mathbf{b}}_q/q\hat{\sigma}^2$: $F$-statistic for testing $H : \tilde{\boldsymbol{\theta}}_q = \mathbf{0}$.
$\hat{b}_0 = \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\theta}}$: estimated intercept.
$\text{cov}(\hat{b}_0, \hat{\boldsymbol{\theta}}) = -(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}}'\hat{\sigma}^2$: estimated vector of covariances of $\hat{b}_0$ with $\hat{\boldsymbol{\theta}}$.
$\hat{v}(\hat{b}_0) = [1/N + \bar{\mathbf{x}}'(\mathcal{X}'\mathcal{X})^{-1}\bar{\mathbf{x}}]\hat{\sigma}^2$: estimated variance of $\hat{b}_0$.
$t_0 = \hat{b}_0/\sqrt{\hat{v}(\hat{b}_0)}$: $t$-statistic, on $N - r$ degrees of freedom for testing hypothesis $H : b_0 = 0$.
$\hat{b}_0 \pm t_{N-r,\frac{1}{2}\alpha}\sqrt{\hat{v}(\hat{b}_0)}$: symmetric $100(1 - \alpha)\%$ confidence interval for $b_0$.

*No-intercept model.* Modify the above expressions as follows.
Use $\mathbf{X}_1$in place of $\mathcal{X}$ :

$\mathbf{X}_1'\mathbf{X}_1$ = matrix of uncorrected sums of squares and products of observed $x$'s.
$\mathbf{X}_1'\mathbf{y}$ = vector of uncorrected sums of products of observed $x$'s and $y$'s.
Put $r = k$ (instead of $k + 1$).
Use $\text{SST} = \mathbf{y}'\mathbf{y}$(instead of $\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$).
Ignore $b_0$ and $\hat{b}_0$.

## 9. EXERCISES

**1** For the following data

| $i$: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 12 | 32 | 36 | 18 | 17 | 20 | 21 | 40 | 30 | 24 |
| $x_i$ | 65 | 43 | 44 | 59 | 60 | 50 | 52 | 38 | 42 | 40 |

with the summary statistics

$$\sum_{i=1}^{10} x_i = 493, \sum_{i=1}^{10} x_i^2 = 25103, \sum_{i=1}^{10} y_i = 250, \sum_{i=1}^{10} x_i y_i = 11654, \sum_{i=1}^{10} y_i^2 = 6994$$

**(a)** Write the normal equations (11)

**(b)** Calculate $\hat{b}$ and $\hat{a}$ as in (14) and (15).

**(c)** Find $\mathrm{SST_m}$, $\mathrm{SSR_m}$ and SSE.

**(d)** Find the coefficient of determination.

**(e)** Make the analysis of variance table and determine whether the regression is statistically significant.

**2** For the growth rate data in Example 19, given that

$$(\mathcal{X}'\mathcal{X})^{-1} = \begin{bmatrix} 0.0966 & -0.0688 & -0.0885 & 0.0457 \\ -0.0688 & 0.9566 & -0.5257 & -0.0880 \\ -0.0885 & -0.5257 & 0.5271 & -0.0541 \\ 0.0457 & -0.0880 & -0.0541 & 0.0769 \end{bmatrix}$$

**(a)** Find a 95% confidence interval on the regression coefficients for France and Germany.

**(b)** Find a 95% confidence interval for the difference in the regression coefficients between France and Germany, that is, $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2$.

**3** **(a)** Show that if, in Example16, we do not consider lack of fit, the regression equation is

$y = 0.17429x - 10.132$ and that the analysis of variance table is

| Source | d.f. | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Model | 1 | 92.9335 | 92.9335 | 67.8347 |
| Error | 18 | 31.2664 | 1.7370 | |
| Total | 19 | 124.2 | | |

**(b)** Find a 95% prediction interval on the predicted value when $x = 75$.

**4** Suppose

$$\hat{\sigma}^2 = 200 \quad \text{and} \quad \hat{\mathbf{b}}' = \begin{bmatrix} 3 & 5 & 2 \end{bmatrix} \text{ where}$$

$$\hat{v}(\hat{b}_1) = 28 \qquad \hat{v}(\hat{b}_2) = 24 \qquad v(\hat{b}_3) = 18$$

$$\hat{\mathrm{cov}}(\hat{b}_1, \hat{b}_2) = -16 \quad \hat{\mathrm{cov}}(\hat{b}_1, \hat{b}_3) = 14 \quad \hat{\mathrm{cov}}(\hat{b}_2, \hat{b}_3) = -12.$$

Show that the $F$-statistic for testing the hypothesis $b_1 = b_2 + 4 = b_3 + 7$ has a value of 1. Calculate the estimate of $\mathbf{b}$ under the null hypothesis.

**5** Show that if, in Example 18, the reduced model is derived by replacing $b_2$ by $b_1 - 4$ the analysis of variance is as follows:

| Source | Degrees of Freedom | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Reduced model | 2 | 1156.7 | 578.35 | 8.49 |
| Error | 3 | 204.3 | 68.1 | |
| Total | 5 | | | |

Is the reduced model statistically significant at $\alpha = 0.05$ or $\alpha = 0.1$?

**6** Since SSM $= \mathbf{y}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{y}$, show that $N^{-1}\mathbf{1}\mathbf{1}'$ is idempotent and that its product with $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'$ is $\mathbf{0}$. What are the consequences of these properties of $N^{-1}\mathbf{1}\mathbf{1}'$?

**7** Derive the matrix $\mathbf{Q}$ such that $\text{SSR}_\text{m} = \mathbf{y}'\mathbf{Q}\mathbf{y}$. Show that $\mathbf{Q}$ is idempotent and that its product with $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the zero matrix. What are the consequences of these properties of $\mathbf{Q}$? Show that $\text{SSR}_\text{m}$ and SSM are independent.

**8** When $\mathbf{y}$ has the variance covariance matrix $\mathbf{V}$, prove that the covariance of the b.l.u.e.'s of $\mathbf{p}'\mathbf{b}$ and $\mathbf{q}'\mathbf{b}$ is $\mathbf{p}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{q}$.

**9** Prove that the definitions in (92) and (93) are equivalent to the computing formula given in (91).

**10** Prove the following results for $\hat{\mathbf{e}}$ for an intercept model. What are the analogous results in a no-intercept model?

**(a)** $\text{cov}(\hat{\mathbf{e}}, \mathbf{y}) = \mathbf{P}\sigma^2$ and $\text{cov}(\hat{\mathbf{e}}, \hat{\mathbf{y}}) = \mathbf{0}_{N\times N}$;

**(b)** $\text{cov}(\hat{\mathbf{e}}, \hat{\mathbf{b}}) = \mathbf{0}_{N\times(k+1)}$ but $\text{cov}(\mathbf{e}, \hat{\mathbf{b}}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$;

**(c)** $\sum_{i=1}^{N} \hat{e}_i y_i = \text{SSE}$ and $\sum_{i=1}^{N} \hat{e}_i \hat{y}_i = 0$.

**11** When $k = 1$ show that (41) and (42) are equivalent to (14) and (15) and also equivalent to (21).

**12** Show that the $F$-statistic for testing the hypothesis $\mathbf{L}'\boldsymbol{\beta} = \mathbf{m}$ takes essentially the same form as $F(H)$. Derive the estimator of $\mathbf{b}$ under the null hypothesis $\mathbf{L}'\boldsymbol{\beta} = \mathbf{m}$.

Show that $\tilde{b}_0 = \hat{b}_0 + \bar{\mathbf{x}}'(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_c)$.

**13** Show that the non-centrality parameters of the non-central $\chi^2$-distribution of SSM, $\text{SSR}_\text{m}$, and SSE add up to SST.

**14** Using the notation of this chapter, derive the $F$-statistics and the values of $\hat{\mathbf{b}}_c$ shown below. In each case, state the distribution of the $F$-statistic under the null hypothesis.

| | Hypothesis | $F$-Statistic | $\tilde{\mathbf{b}}$ |
|---|---|---|---|
| (i) | $\boldsymbol{\mathscr{C}} = \mathbf{0}$ | $\dfrac{\text{SSR}_{\text{m}}}{k\hat{\sigma}^2}$ | $\tilde{\mathbf{b}}' = \begin{bmatrix} \bar{y} & \mathbf{0}' \end{bmatrix}$ |
| (ii) | $\boldsymbol{\mathscr{C}} = \mathbf{b}_0$ | $\dfrac{(\hat{\mathbf{b}} - \boldsymbol{\mathscr{C}}_0)'\mathcal{X}'\mathcal{X}(\hat{\mathbf{b}} - \boldsymbol{\mathscr{C}}_0)}{k\hat{\sigma}^2}$ | $\tilde{\mathbf{b}}' = \begin{bmatrix} \bar{y} - \bar{x}'\boldsymbol{\mathscr{C}}_0 & \boldsymbol{\mathscr{C}}_0' \end{bmatrix}$ |
| (iii) | $\lambda'\boldsymbol{\mathscr{C}} = \mathbf{m}$ | $\dfrac{(\lambda'\hat{\boldsymbol{\mathscr{C}}} - \mathbf{m})^2}{\lambda'(\mathcal{X}'\mathcal{X})^{-1}\lambda\hat{\sigma}^2}$ | $\tilde{\mathbf{b}}' = \hat{\mathbf{b}} + \left(\dfrac{\lambda'\hat{\boldsymbol{\mathscr{C}}} - \mathbf{m}}{\lambda'(\mathcal{X}'\mathcal{X})^{-1}\lambda}\right)\begin{bmatrix} \bar{\mathbf{x}}' \\ -\mathbf{I} \end{bmatrix}(\mathcal{X}'\mathcal{X})^{-1}\lambda$ |
| (iv) | $\boldsymbol{\mathscr{C}}_q = \mathbf{0}$ | $\dfrac{\hat{\mathbf{b}}_q \tilde{\mathscr{T}}_{qq} \hat{\mathbf{b}}_q}{q\hat{\sigma}^2}$ | $\tilde{\mathbf{b}}' = \begin{bmatrix} \bar{y} - \bar{\mathbf{x}}_p'\hat{\boldsymbol{\mathscr{C}}}_p + \bar{\mathbf{x}}_p'\tilde{\mathscr{T}}_{pq}\tilde{\mathscr{T}}_{qq}^{-1}\hat{\mathbf{b}}_q \\ 0 \\ \hat{\boldsymbol{\mathscr{C}}}_p - \tilde{\mathscr{T}}_{pq}\tilde{\mathscr{T}}_{qq}^{-1}\hat{\mathbf{b}}_q \end{bmatrix}$ |

**15** **(a)** By using Expression (131) prove directly that $[\mathbf{y}'\mathbf{y} - (\text{SSE} + \text{Q})]/\sigma^2$ has a non-central $\chi^2$-distribution, independent of SSE when $\mathbf{m} = \mathbf{0}$.

**(b)** Show that under the null hypothesis, the non-centrality parameter is $\mathbf{b}'\mathbf{L}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})\mathbf{L}'\mathbf{b}/2\sigma^2$.

**16** Prove that in Table 3.7 $\text{SSR}-\text{Q} = \hat{\mathbf{b}}_{pc}'\mathbf{x}_p'\mathbf{X}_p\hat{\mathbf{b}}_{pc}$. *Hint:* Use (132), $(\mathbf{X}'\mathbf{X})^{-1}$ defined before (122) and the formula for the inverse of a partitioned matrix.

**17** If $\hat{b}_{k+1}$ is the estimated regression coefficient for the $(k+1)$th independent variable in a model having just $k+1$ such variables, the corresponding $t$-statistic for testing the hypothesis $b_{k+1} = 0$ is $t = \hat{b}_{k+1}/\sqrt{\hat{\text{var}}(\hat{b}_{k+1})}$ where $\hat{\text{var}}(\hat{b}_{k+1})$ is the estimated variance of $\hat{b}_{k+1}$. Prove that the $F$-statistic for testing the same hypothesis is identical to $t^2$.

**18** Assume $\mathbf{X}$ is of full rank. For $\boldsymbol{\lambda}'$ and $\hat{\mathbf{b}}$ in (48) and (49), $\mathbf{t}'\hat{\mathbf{b}} = \boldsymbol{\lambda}'\mathbf{y}$ is the unique b.l.u.e. of $\mathbf{t}'\mathbf{b}$. Prove this by assuming that $\mathbf{t}'\hat{\mathbf{b}} + \mathbf{q}'\mathbf{y}$ is a b.l.u.e. different from $\mathbf{t}'\hat{\mathbf{b}}$ and showing that $\mathbf{q}'$ is null.

**19** Consider the linear model

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_5 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_5 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_5 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + e$$

where

$$\mathbf{y}' = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & y_{21} & y_{22} & y_{23} & y_{24} & y_{25} & y_{31} & y_{32} & y_{33} & y_{34} & y_{35} \end{bmatrix}$$

**(a)** Show that $b_i = \bar{y}_{i.}, i = 1, 2, 3$.

**(b)** Show that $\text{SSR}_{\text{m}} = \frac{y_{1.}^2}{5} + \frac{y_{2.}^2}{5} + \frac{y_{3.}^2}{5} - \frac{y_{..}^2}{15}$

**(c)** For each of the hypothesis $H : b_1 - b_2 = 0, H : b_1 + b_2 - 2b_3 = 0$ find:

**1.** The $F$-statistic.

**2.** $\hat{\mathbf{b}}_c$.

**20** Show how to establish the algebraic equivalence of the two expressions in (53) making use of the matrix identity (57).

**21 (a)** Assume that the assumptions on $\mathbf{b}$ in Section 3e with $\mathbf{V} = \sigma^2\mathbf{I}$ hold true. Show that the estimator of the form $\hat{\mathbf{b}}^{(b)} = a + \mathbf{c}'\hat{\mathbf{b}}$ that minimizes

$$v = Var(\mathbf{p}'\mathbf{b} - a - \mathbf{c}'\hat{\mathbf{b}})$$

subject to

$$E(\mathbf{p}'\mathbf{b} - a - \mathbf{c}'\hat{\mathbf{b}}) = 0$$

is

$$\mathbf{p}'\hat{\mathbf{b}}^{(b)} = \mathbf{p}'\theta + \mathbf{p}'\mathbf{F}[\mathbf{F} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\hat{\mathbf{b}} - \theta).$$

**(b)** Show that the estimator in (a) is algebraically equivalent to (53) when $\mathbf{V} = \sigma^2\mathbf{I}$.

**(c)** Show that the minimum variance for the estimator derived in (a) is

$$v = \mathbf{p}'\mathbf{F}\mathbf{p} - \mathbf{p}'\mathbf{F}[\mathbf{F} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{F}\mathbf{p}.$$

**22** Show that when in 21a

$$\theta = 0, \mathbf{F} = \frac{\sigma^2}{(1-d)k}[\mathbf{I} + dk(\mathbf{X}'\mathbf{X})^{-1}], 0 < d < 1,$$

$$\mathbf{p}'\hat{\mathbf{b}}^{(b)} = d\mathbf{p}'\hat{\mathbf{b}} + (1-d)\mathbf{p}'[\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{y}.$$

# 4

# INTRODUCING LINEAR MODELS: REGRESSION ON DUMMY VARIABLES

This chapter begins by describing, in terms of an example, a type of regression analysis that is not recommended. However, it gives a motivation for the use of an alternative analysis already mentioned in Chapter 0, regression on dummy (0, 1) variables. Sections 1a, 1b, and 1c will propose other coding methods and explain their disadvantages as compared to the use of dummy variables. Dummy (0, 1) variables will be very useful for the study on non-full-rank linear models, the subject of Chapter 5.

## 1. REGRESSION ON ALLOCATED CODES

### a. Allocated Codes

The Bureau of Labor Statistics Consumer Survey July 2013–June 2014 reports detailed data about household expenditure habits and the characteristics of each household sampled. One of many questions that could be asked about such data is, "To what extent is a household's investment in consumer durables associated with the occupation of the head of the household?" Investment behavior is, of course, related to many factors other than occupation. However, for purposes of illustration, we consider this question just as it stands.

The survey data contain figures on investment in consumer durables (hereafter simply referred to as investment) for some 126,420,000 consumer units. A consumer

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

unit is either a family or a single consumer. In addition, for each unit, the occupation of the principal breadwinner is recorded, in one of four categories. These are

1. Laborer
2. Skilled technical
3. Professional, manager, or service
4. Self-employed

Suppose that a regression has been proposed of investment on occupation, as a means of answering the question posed in quotes above? The problem that immediately arises is that of how can occupation be measured. One possibility is to measure it by the code numbers 1, 2, 3, and 4 listed above. In some sense, one might rationalize that these numbers correspond to a measure of occupational status. One might also ask how else can one "measure" occupation recorded in this way in order to investigate the effect of occupation on investment. Accepting these numbers 1, 2, 3, and 4, the procedure would be to carry out a regression analysis of $y$, investment on $x$, which would be 1, 2, 3, or 4 depending on which occupational category the principal breadwinner belonged to. Details of the regression analysis would proceed in the usual fashion using a model

$$E(y_i) = b_0 + b_1 x_i. \tag{1}$$

A test of the hypothesis $b_1 = 0$ could easily be made.

### b. Difficulties and Criticism

As an analysis procedure, what we have just described is okay. However, an inherent difficulty occurs with the definition of $x$, the independent variable occupational status. Although the four categories of occupation present different kinds of occupation, allocation of the numbers 1, 2, 3, and 4 to these categories as "measures of occupational status" may not accurately correspond to the underlying measure of whatever is meant by occupational status. The allocation of the numbers is in this sense, quite arbitrary. For example, does a professional man have three times as much status as a laborer? If the answer is "no" and a different set of numbers is allocated to the categories, the same kind of criticism may be leveled. In fact, whatever the allocation is, it is essentially arbitrary.

This allocation of codes causes problems relative to the suggested models (1). By giving a self-employed person an $x$-value of 4, we are not really saying he has twice as much status as an artisan (for whom $x = 2$). What we are saying in terms of the model is that

$$\begin{aligned} E(\text{investment of a laborer}) &= b_0 + b_1, \\ E(\text{investment of a skilled worker}) &= b_0 + 2b_1, \\ E(\text{investment of a professional}) &= b_0 + 3b_1 \end{aligned}$$

and

$$E(\text{investment of a self-employed}) = b_0 + 4b_1.$$

This means, for example, that

$$\begin{aligned} &E(\text{investment of a self-employed}) - E(\text{investment of an skilled worker}) \\ &\quad = E(\text{investment of a professional}) - E(\text{investment of a laborer}) \\ &\quad = 2[E(\text{investment of a professional}) - E(\text{investment of a skilled worker}) \\ &\quad = 2b_1. \end{aligned} \tag{2}$$

This, in terms of the real world, may be quite unrealistic. Yet, even without data, allocation of the numbers 1, 2, 3, and 4 forces this consequence on the analysis. The only estimation the analysis will yield will be that of $b_1$ and $b_0$. This will also be the case if a set of numbers different from 1, 2, 3, and 4 is allocated to the categories. Relationships akin to (2) will still apply and, so far as they are concerned, estimation of $b_1$ will be the only achievement from the regression analysis.

The inherent difficulty with the analysis suggested above is the allocation of codes to non-quantitative variables such as "occupation." Yet, such variables are frequently of interest. Examples include religion and nationality in the behavioral sciences; species, fertilizer, and soil type in agriculture; source of raw material, treatment, and plant location in an industrial process; and many others. Allocating codes to these variables involves at least two difficulties. First, often it cannot be made a reasonable procedure (e.g., allocating codes to "measure" geographical regions of the United States). Second, by making any such allocation, we automatically impose value differences on the categories of the variables in the manner illustrated in equation (2).

### c. Grouped Variables

The same difficulties also arise with variables that are more measurable than those just considered. Education is an example. It can be measured as the number of years if formal. An immediate question is, when does formal education start? Measurement difficulties of this sort can, of course, be avoided by defining education as a series of categories, such as high school incomplete, high school graduate, college graduate, and advanced degree. These are not unlike the categories of occupation discussed earlier. However, they do have a clear-cut sense of ordinality about them and hence some sense of "measure." This sense of ordinality would disappear at once, upon addition of a fifth category "foreign education." The matter is also complicated by subjectivity of decisions that have to be made in classifying people within such categories. For example, how would a person with a foreign education, but an American doctorate, be classified? What would be the classification of a college dropout who had subsequently passed the Institute of Flycatchers examination?

Many instances could be cited where variables are grouped into categories in a manner similar to the education example just given. Income is a common example, with such categories as high, medium, low, and poor. Another example is city size, such as metropolis, large city, town, and village, and so on. In all these cases and many others, it is possible but for the reasons just described, not very rational to impose codes on the categories of independent variables of this nature. This problem is avoided by the technique of regression on (0, 1) dummy variables. As an analysis procedure, it is also more informative than regression on allocated codes because it leads to a larger multiple correlation coefficient. Recall that the multiple correlation coefficient was defined in Section 4g of Chapter 3. Furthermore, it provides from the data, estimated values to be associated with categories of the independent variables, rather than allocating codes arbitrarily, regardless of the data. Searle and Udell (1970) carried out both regression on allocated codes and regression on dummy variables for the same set of data. By doing this, they were able to give examples of estimated values and larger $R^2$ values that were available using dummy variables.

### d. Unbalanced Data

Despite the limitations of using allocated codes, an investigator with data to analyze with limited training and experience in statistics might well be tempted to use these codes. Armed with the knowledge of regression and of analysis of variance as depicted from the point of view of carefully designed experiments (albeit a good knowledge of these topics), an investigator could easily feel that regression on allocated codes was an appropriate analysis. For example, for 100 people in a hypothetical (pilot) survey designed to investigate the effect of both occupation and education on investment, suppose that the number of people reporting data were distributed as in Table 4.1. Faced with data from people so classified, the choice of an analysis procedure may not be easy for some investigators. A patent difficulty with such data is that the numbers of observations in the subclasses of the data are not all the same. Data where these numbers are the same are known as *equal numbers data* or more frequently *balanced data*. In contrast, those like Table 4.1 with unequal numbers of observations in the subclasses, including perhaps some that contain no observations

**TABLE 4.1 Number of People, Classified According to Occupation and Education, Who Reported Investment Data**

| | Education | | |
|---|---|---|---|
| Occupation | High School Incomplete | High School Graduate | College Graduate |
| Laborer | 14 | 8 | 7 |
| Artisan | 10 | — | — |
| Professional | — | 17 | 22 |
| Self-employed | 3 | 9 | 10 |

at all (empty subclasses, or empty cells), are called *unequal-numbers data* or, more usually, *unbalanced data*, or sometimes "messy" data.

Traditional analysis of variance methods, in terms of well-designed experiments, are generally only applicable to balanced data. (Exceptions are the specified patterns of Latin square designs, balanced incomplete block designs, and derivatives thereof.) Hence, for unbalanced data like those of Table 4.1, analysis of variance in its traditional framework is inapplicable. On the other hand, regression can be used with some degree of propriety by allocating codes to "education" and "occupation." Disadvantages implicit in doing this are incurred. These have been described already. However, at least *some* analysis can be conducted. A computer can do the arithmetic, and interpretation is straightforward. The possibility that regression on allocated codes may be used must therefore not be ignored. Indeed, in the presence of powerful computer programs for regression analysis, the possibility of its being used is greatly increased.

In light of the advantages already discussed, the preferred analysis is regression on dummy (0, 1) variables. Furthermore, this technique is identical to the established analysis of variance procedures that are available for unbalanced data. In addition to being called regression on dummy variables, or analysis of variance of unbalanced data, it is also known as the *method of fitting constants*. This means fitting the constants or the terms of a linear model. For this method of analysis, the computations for unbalanced data are usually more complicated than those of traditional analysis of variance for balanced data. As a result, before the modern computer era, there was a limited demand for the analysis of unbalanced data. Due to the ability of the modern computer to store and process information, there is a greater demand for analysis of unbalanced data that *cannot* be made by minor adjustments to traditional analyses of variance for balanced data. Unbalanced data have their own analysis of variance techniques. Those for balanced data are merely special cases of the analysis of variance methods for unbalanced data. As we shall see, unbalanced data analysis can be couched in matrix expressions, many of which simplify very little in terms of summation formulae. However, when the number of observations in each of the subclasses is the same, the matrix expressions simplify considerably. In fact, they reduce to the well-known summation formulae of traditional analysis of variance of designed experiments. These designed experiments include randomized blocks, factorial experiments, and many others.

Therefore, one can think of analysis of variance for balanced data as special cases of analysis of variance for unbalanced data. This is the attitude taken in this book. In Chapter 5, we develop general analysis procedures. We apply them to specific situations in Chapters 6, 7, and 8 but always for unbalanced data. Reference will be made to simplification of the results for balanced data.

The remainder of this chapter serves as a preface to the general linear model theory in the chapters that follow. Regression on dummy variables is identical for a wide class of linear models. Thus, we can introduce linear models by presenting regression on dummy variables. Despite the widespread use of regression on dummy variables in many fields of applications, its equivalence to linear models is not

always appreciated. As a result, the users of regression on (0, 1) dummy variables do not always adopt ramifications of linear models. Therefore, we plan to do the following:

(i) Present regression on (0, 1) dummy variables;
(ii) Demonstrate the equivalence of regression on (0, 1) dummy variables to linear models;
(iii) Characterize the description of linear models;
(iv) Confine our attention to linear models thereafter.

## 2. REGRESSION ON DUMMY (0, 1) VARIABLES

### a. Factors and Levels

We shall enhance the discussion of regression on dummy variables by making use of the notion of factors and levels. We adapt this useful descriptive terminology from the literature on experimental design.

In studying the effect of the variables "occupation" and "education" on investment behavior, as in Table 4.1, we are interested in the extent to which each category of each variable is associated with investment. Thus, we are interested in seeing to what extent a person's being a skilled worker affects his/her investment and to what extent someone else's being self-employed affects his/her investment. In particular, we are interested in investigating the difference between the effects of these two categories in the population of people of whom our data are considered to be a random sample. The terms "factor" and "level" are introduced to acknowledge the immeasurability of the variables and the associated arbitrariness or subjectivity in deciding their categories as discussed in the previous section. The word *factor* denotes what heretofore was called a variable. Thus, occupation is one factor and education is another. The categories into which each factor (variable) has been divided are called *levels* of that factor. Thus, laborer is one level of the factor occupation and professional is another level of that factor. There are four levels of the factor occupation and three levels of the factor education (see Table 4.1). The reason the term "factor" is used instead of variable is that a factor cannot be measured by cardinal values whereas a variable can be. We reserve the term "variable" for items that can be measured by cardinal values. Given this interpretation of the term "variable," the only variable in our investigation is investment. Other elements of the investigation are factors, each with a number of levels. The term "levels" emphasizes that the groupings of a factor are just arbitrary divisions with no imposition on allocated values. It is these that we seek to estimate from data. In this context, the ordinal numbers 1, 2, 3, and 4 shown in the list of observations are no longer values given to the category of a variable. They are used solely to identify levels of factors. For example, level 2 of the occupation factor is skilled worker.

Thinking in terms of levels of factors rather than groupings of variables overcomes many of the difficulties inherent in using allocated codes. Even when groupings of a non-quantitative variable have no sense of ordinality, they can still be thought about as levels of a factor. Whenever value differences cannot be placed rationally on the groupings, the concept of levels enables us to estimate differences between the effects that the levels of the factor have on the variable being studied, without any a priori imposition of values. This estimation of differences is brought about by regression on dummy (0, 1) variables.

### b. The Regression

Our aim is to consider the effects of the levels of each factor on investment. We begin by estimating just the effect of education on investment. In particular, we estimate the effect on investment of each of the three levels of the factor education shown in Table 4.1. To accomplish this, we set up a regression on three independent variables $x_1, x_2$, and $x_3$:

$$y_i = b_0 + b_{i1}x_1 + b_{i2}x_2 + b_{i3}x_3 + e_i. \tag{3}$$

In this context, $y_i$ is investment, and $b_0$ and $e_i$ are respectively the constant and error terms found in regression analysis. Corresponding to the independent variables, the $x$'s yet to be defined are the regression coefficients $b_1, b_2$, and $b_3$. As a result of how the $x$'s will be defined, these $b$'s will turn out to be terms that lead to estimates of the differences between the effects of investment on the levels of the factor education.

To define the $x$'s, note that each person falls into one and only one educational level. For whichever level he/she is in, let the corresponding $x$ take the value unity. Let all other $x$'s for that person have a value of zero. Thus, a high school graduate is in level 2 of the education factor. For him or her, we have that $x_{i2} = 1$ with $x_{i1} = 0$ and $x_{i3} = 0$. In this way, the numerical values (0's and 1's) can be assigned to all three $x$'s for each person in the data. A regression analysis is carried out on these values.

It is because each $x$-value is unity when someone belongs to the corresponding level of education, and zero otherwise, that the $x$'s are described as (0, 1) variables. They are called "dummy variables" because they are not true variables in the sense previously defined. Despite this, the formal procedures of regression can be carried out, with consequences of great interest.

**Example 1 Linear Models Representation with Dummy Variables** Assume that we have investment data on three people who did not finish high school, on two who did, and one college graduate. These six observations (investment indices) are shown in Table 4.2, where $y_{ij}$ is the observation on the $j$th person in the $i$th level of educational status. Then with $e_{ij} = y_{ij} - E(y_{ij})$, just as in regression (except for

**TABLE 4.2 Investment Indices of Six People**

| | Educational Status | Investment Index |
|---|---|---|
| 1 | (High school incomplete) | $y_{11}, y_{12}, y_{13}$ |
| 2 | (High school graduate) | $y_{21}, y_{22}$ |
| 3 | (College graduate) | $y_{31}$ |

having two subscripts rather than one), we write the observations in terms of (3) as follows:

$$\begin{aligned}
y_{11} &= b_0 + b_1(1) + b_2(0) + b_3(0) + e_{11} \\
y_{12} &= b_0 + b_1(1) + b_2(0) + b_3(0) + e_{12} \\
y_{13} &= b_0 + b_1(1) + b_2(0) + b_3(0) + e_{13} \\
y_{21} &= b_0 + b_1(0) + b_2(1) + b_3(0) + e_{21} \\
y_{22} &= b_0 + b_1(0) + b_2(1) + b_3(0) + e_{22} \\
y_{31} &= b_0 + b_1(0) + b_2(0) + b_3(1) + e_{31}.
\end{aligned}$$

The 1's and 0's in parentheses are the values of the dummy (0, 1) variables. Their pattern can be seen more clearly when the equations are written in matrix form as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \end{bmatrix} \tag{4}$$

By writing

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \tag{5}$$

the equations become the familiar form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{6}$$

that has been dealt with so fully in Chapter 3. □

There are some noteworthy things about the model (6) in Example 1 above. If we define the properties of the $\mathbf{e}$ term exactly as in regression, namely $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$,

application of least squares to model (6) yields the same normal equations as before, $\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}$. However, $\mathbf{X}$ does not have full-column rank. As can be seen in (5) the sum of the last three columns equals the first. This model is described as a "*model not of full rank*". Its property is that $\mathbf{X}$ does not have full-column rank. The important consequence is that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. As a result, $\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}$ cannot be solved as $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Solutions can be found by using a generalized inverse $\mathbf{G}$ in place of $\mathbf{X}'\mathbf{X}$, that is, $\hat{\mathbf{b}} = \mathbf{G}\mathbf{X}'\mathbf{y}$. We will discuss these solutions in Chapter 5. Before doing this, we shall give another example and discuss other aspects of linear models.

**Example 2 Another Linear Model with Dummy Variables** Countless experiments are undertaken each year in agriculture and the plant sciences to investigate the effect of growth and yield of various fertilizer treatments applied to different varieties of a species. Suppose we have data from six plants, representing three varieties being tested in combination with two fertilizer treatments. Although the experiment would not necessarily be conducted by growing the plants in varietal rows, it is convenient to visualize the data in Table 4.3. The entries of the table are such that $y_{ijk}$ represents the yield of the $k$th plant of the variety $i$ that received the treatment $j$.

We will now write these out, using five dummy (0, 1) variables and five regression coefficients corresponding to the three varieties and two treatments. The regression coefficients for the three varieties will be denoted by $\alpha_1, \alpha_2$, and $\alpha_3$ and those for the treatments will be $\beta_1$ and $\beta_2$. Furthermore, the intercept term in the regression, previously denoted by $b_0$ will now be written as $\mu$. Thus, the vector of parameters will be

$$\mathbf{b}' = [\,\mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \beta_1 \quad \beta_2\,].$$

This notation clearly distinguishes between regression coefficients for varieties ($\alpha$'s) and those for treatments ($\beta$'s). In contrast to using $b$'s as elements of **b,** it avoids double subscripting which could then provide that clarity. With this notation, the regression equation for $y_{ijk}$ is

$$y_{ijk} = \mu + \alpha_1 x_{ijk,1} + \alpha_2 x_{ijk,2} + \alpha_3 x_{ijk,3} + \beta_1 x^*_{ijk,1} + \beta_2 x^*_{ijk,2} + e_{ij}$$

**TABLE 4.3 Yields of Six Plants**

| | Treatment | |
|---|---|---|
| Variety | 1 | 2 |
| 1 | $y_{111}, y_{112}$ | $y_{121}$ |
| 2 | $y_{211}$ | $y_{221}$ |
| 3 | $y_{311}$ | |

where the $x$'s and the $x^*$'s are the dummy variables. Thus, the regression equations for the yields in Table 4.3 are

$$\begin{aligned}
y_{111} &= \mu + \alpha_1(1) + \alpha_2(0) + \alpha_3(0) + \beta_1(1) + \beta_2(0) + e_{111} \\
y_{112} &= \mu + \alpha_1(1) + \alpha_2(0) + \alpha_3(0) + \beta_1(1) + \beta_2(0) + e_{112} \\
y_{121} &= \mu + \alpha_1(1) + \alpha_2(0) + \alpha_3(0) + \beta_1(0) + \beta_2(1) + e_{121} \\
y_{211} &= \mu + \alpha_1(0) + \alpha_2(1) + \alpha_3(0) + \beta_1(1) + \beta_2(0) + e_{211} \\
y_{221} &= \mu + \alpha_1(0) + \alpha_2(1) + \alpha_3(0) + \beta_1(0) + \beta_2(1) + e_{221} \\
y_{311} &= \mu + \alpha_1(0) + \alpha_2(0) + \alpha_3(1) + \beta_1(1) + \beta_2(0) + e_{311}.
\end{aligned} \tag{7}$$

Using $\mathbf{y}$ and $\mathbf{e}$ to denote the vectors of observations and error terms in the usual way, these equations become

$$\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{e}. \tag{8}$$

Now write

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix}. \tag{9}$$

The matrix $\mathbf{X}$ is not of full column rank. The sum of columns 2, 3, and 4 equals that of column 1, as does that of columns 5 and 6. With this proviso, equations (8) is $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ just as before, the equation of a model that is not of full rank. □

In general, the matrix $\mathbf{X}$, having elements that are all 0 or 1, is called an *incidence matrix*, because the presence of the 1's among the elements describes the incidence of the terms in the model ($\mu$, the $\alpha$'s and the $\beta$'s) in the data.

## 3. DESCRIBING LINEAR MODELS

### a. A One-Way Classification

Equations (4) and (8) in Examples 1 and 2 of Section 2b have been developed from the point of view of regression on dummy (0, 1) variables. Consider equations (4)

again. They relate to investment indices of six people in three different levels of educational status, as shown in Table 4.2. Suppose the equations (4) are rewritten as

$$\begin{aligned} y_{11} &= \mu + b_1 + e_{11}, \\ y_{12} &= \mu + b_1 + e_{12}, \\ y_{13} &= \mu + b_1 + e_{13}, \\ y_{21} &= \mu + b_2 + e_{21}, \\ y_{22} &= \mu + b_2 + e_{22} \end{aligned} \tag{10}$$

and

$$y_{31} = \mu + b_3 + e_{31},$$

where the $x$'s are no longer shown explicitly and $\mu$ is written for $b_0$. Observe that in each equation of (10) the subscript of the $b$ corresponds exactly to the first subscript of the $y$. For example, $b_1$ is found $y_{11}, y_{12}$, and $y_{13}$ and $b_2$ is in $y_{21}$ and $y_{22}$. Hence, each equation of (10) can be written as

$$y_{ij} = \mu + b_i + e_{ij} \tag{11}$$

for the various values that $i$ and $j$ take in the data. In this case $i = 1, 2, 3$ and the upper limit on $j$ in the $i$th class is the number of observations in that $i$th class. Denoting this by $n_i$, we have $j = 1, 2, \ldots, n_i$ where $n_1 = 3, n_2 = 2$ and $n_3 = 1$. Thus, we have developed (11) as the equation of the general linear model for three classes. For $a$ classes it applies for $i = 1, 2, \ldots, a$.

Although (11) is the general form of a linear model equation, its specific values are still as shown in (4), exactly as developed in the regression context. Now, however, there is no need to view the elements of **b** as regression coefficients, nor the 0's and 1's of **X** as dummy variables. The elements of **b** can be given meanings in their own rights and the 0's and 1's of **X** relate to "absence" and "presence" of levels of factors.

Since $\mu$ enters into every element of (10), it is described as the general mean of the population of investment indices. It represents some overall mean regardless of educational status.

To give meaning to the $b$'s consider $b_1$. In equations (10) (or (4), they are equivalent) $b_1$ occurs only in those equations pertaining to investment indices of people of educational status (high school incomplete) namely $y_{11}, y_{12}$, and $y_{13}$. Likewise, $b_2$ occurs only in the equations for people of educational status 2 (high school graduate), $y_{21}$ and $y_{22}$. Likewise, $b_3$ is in the equation for $y_{31}$ and nowhere else. Thus, $b_1$ gets described as the effect on investment of a person being in educational status 1. Similar descriptions apply to $b_2$ and $b_3$. In general, in terms of (11), $b_i$ is described as the *effect* on investment due to educational status $i$.

Description of a linear model is incomplete without specifying distributional properties of the error terms, the $e_{ij}$'s evident in equations (4), (10), and (11). Usually, this is done by attributing them with the same kinds of properties as in regression

analysis (see equations (5), (6), and (7) in Section 1b of Chapter 3). Thus, $e_{ij}$ is defined as $e_{ij} = y_{ij} - E(y_{ij})$.

Then $E(e_{ij}) = 0$, giving

$$E(y_{ij}) = \mu + b_i.$$

The variance of each $e_{ij}$ is defined as $\sigma^2$ and so

$$v(e_{ij}) = E[e_{ij} - E(e_{ij})]^2 = \sigma^2, \quad \text{for} \quad \text{all } i \text{ and } j.$$

Furthermore, covariances between all pairs of different $e$'s are taken to be zero. Thus,

$$\text{cov}(e_{ij}, e_{i'j'}) = 0 \text{ unless } i = i' \text{ and } j = j'$$

in which case the covariance becomes the variance $\sigma^2$. As a result,

$$\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}.$$

The general description of the one-way classification model can therefore be summarized as follows. For $y_{ij}$ being the $j$th observation in the $i$th class, the equation of the model is (11) given by

$$y_{ij} = \mu + b_i + e_{ij}.$$

The term $\mu$ is the general mean, $b_i$ is the effect on $y_{ij}$ due to the $i$th class and $e_{ij}$ is a random error term peculiar to $y_{ij}$ with

$$\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I}).$$

For $a$ classes, we have that $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n_i$ for the $i$th class. The additional assumption of normality is made when hypothesis testing and confidence intervals are considered. We then assume that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

### b. A Two-Way Classification

Suppose equations (7) are rewritten with the $x$'s no longer explicitly shown, just as were equations (4) from Example 1 in (10). Then (7) becomes

$$\begin{aligned}
y_{111} &= \mu + \alpha_1 + \beta_1 + e_{111} \\
y_{112} &= \mu + \alpha_1 + \beta_1 + e_{112} \\
y_{121} &= \mu + \alpha_1 + \beta_2 + e_{121} \\
y_{211} &= \mu + \alpha_2 + \beta_1 + e_{211} \\
y_{221} &= \mu + \alpha_2 + \beta_2 + e_{221} \\
y_{311} &= \mu + \alpha_3 + \beta_1 + e_{311}
\end{aligned} \tag{12}$$

Here, in each equation, the subscripts on the $\alpha$ and $\beta$ correspond to the first two on $y$. Notice that $\alpha_1$ and $\beta_1$ are found in $y_{111}$ and $y_{112}$, and $\alpha_2$ and $\beta_1$ are in $y_{211}$. Hence, each equation in (12) can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}. \tag{13}$$

The values taken by $i$, $j$, and $k$ in the data are in this case, $i = 1, 2, 3$ with the upper limit of $k$ being the number of observations in the $i$th variety receiving the $j$th treatment. Denoting this by $n_{ij}$, we have $k = 1, 2, \ldots, n_{ij}$, where $n_{11} = 2, n_{12} = 1$, $n_{22} = 1, n_{31} = 1$, and $n_{32} = 0$. Thus, (13) is the equation of the general linear model involving varieties and treatments.

As with the one-way classification of the preceding section, so here, the elements of $\mathbf{b}$ (in this case $\mu$, the $\alpha$'s and $\beta$'s) do not need to be viewed as regression coefficients but can be given meanings in their own rights. First, $\mu$ is described as the mean of the whole population of yields, representing some overall mean regardless of variety or treatment. Second, in (12) or equivalently (7), $\alpha_1$ occurs only in those equations pertaining to yields of variety 1, namely $y_{111}, y_{112}$, and $y_{121}$. Likewise, $\alpha_2$ occurs only in those equations of yields of variety 2, $y_{211}$ and $y_{221}$. Similarly, $\alpha_3$ is in the equation for $y_{311}$ and nowhere else. Thus, $\alpha_1$ gets described as the effect on yield of a plant being of variety 1. Similar descriptions apply to $\alpha_2$ and $\alpha_3$. In general, $\alpha_i$ is described as the effect due to variety $i$.

For the $\beta$'s, $\beta_1$ occurs only in equations in yields that received treatment 1, $y_{111}, y_{112}, y_{211}$, and $y_{311}$ and $\beta_2$ is only in the equations pertaining to treatment 2, those for $y_{121}$ and $y_{221}$. Thus, $\beta_j$ is described as the effect due to treatment $j$. Hence, general description of the $\beta$'s is similar to that of the $\alpha$'s. Both are effects on yield. However, the $\alpha$'s are effects due to variety while the $\beta$'s are effects due to treatments.

The error terms in this model, the $e_{ijk}$, are assumed to have the same properties as before. If $\mathbf{e}$ is the vector of the $e_{ijk}$, we assume that $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. For formulating hypothesis tests and confidence intervals, we also assume normality of the error terms.

Apart from $\mu$ and $e_{ijk}$, equation (13) has terms for just two factors. These can be referred to as an $\alpha$-factor and a $\beta$-factor. The model with equation (13) could be called a two-factor model, although the name two-way classification is more firmly established.

Its general description is as follows. For $y_{ijk}$ being the $k$th observation on the $i$th level of the $\alpha$-factor and the $j$th level of the $\beta$-factor, the equation of the model is (13):

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}.$$

The general mean is $\mu$. The effect on $y_{ijk}$ due to the $i$th level of the $\alpha$-factor is $\alpha_i$. The effect on $y_{ijk}$ due to the $j$th level of the $\beta$-factor is $\beta_j$. The random error term peculiar to $y_{ijk}$ is $e_{ijk}$ with $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$.

Of course, for hypothesis testing and confidence intervals we assume

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

The example of this model is from agriculture. Historically, the analysis of variance that we will study in the succeeding chapters was developed for agriculture. However, it has found application in many other fields of human endeavor. There can be many applications that make use of the same statistical methodology. Indeed the two-factor model we have just formulated can apply to many other situations. For the example of Table 4.1 concerning the effect of occupation and education on investment, equation (13) could act equally as well as it could for the agricultural example. The $\alpha_i$ would then be the effect on investment of the $i$th occupation category (the $i$th level of the occupation factor). The $\beta_j$ would be the effect of the $j$th level of the education factor.

Similarities between the above description of the two-way classification and the one-way classification at the end of Section 3 a will be clearly apparent. They extend quite naturally to many-factored models. The following outline of the three-way classification illustrates this.

### c. A Three-Way Classification

Suppose that for the data of Table 4.1 the hometown region of the United States (Northeast, South, Midwest, Southwest, Rockies or West Coast) was also recorded for each person. A study of the effects on investment of occupation, education, and region could then be made using a model whose equation is

$$y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijkh} \tag{14}$$

In equation (14), the terms have the following interpretations. The $y_{ijkh}$ is the investment index of the $h$th person in the $i$th occupation and the $j$th level of education in the $k$th region. The general mean is $\mu$. The effect on investment due to the $i$th occupation is $\alpha_i$.

The $\beta_j$ is the effect due to the $j$th level of education. The effect due to the $k$th region is $\gamma_k$. As before, $e_{ijkh}$ is an error term peculiar to $y_{ijkh}$. Again, we assume that $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. If in the data there are $a$ levels of occupation, then $i = 1, 2, \ldots, a$. If there are $b$ levels of education, then $j = 1, 2, \ldots, b$. For $c$ regions, we have $k = 1, 2, \ldots, c$. For $n_{ijk}$ observations, in the subclass of the data representing the $i$th occupation, the $j$th level of education and the $k$th region, we have $h = 1, 2, \ldots, n_{ijk}$.

Extension of models of this nature to four-way and higher order classifications should be clear.

### d. Main Effects and Interactions

**(*i*) *Main Effects.*** The $\alpha$'s, $\beta$'s, and $\gamma$'s of Examples 1 and 2 each represent the effect of $y$ of one level on one factor. Thus, in the two-way classification of Table 4.3, $\alpha_i$ of equation (13) refers to the yield of the $i$th level of the factor variety, specifically

variety $i$. Likewise, $\beta_j$ in the same equation refers to the yield of treatment $j$. Effects of this nature that pertain to a single level of the factor are called *main effects*. This is logical because the effects of variety and treatment on yield are where our main interest lies. Hence, the elements of the model that correspond to them are called the main effects of the model.

By its very nature, the equation of the model implies that the effect $\alpha_i$ is added to the effect $\beta_j$ in conjecturing the value of $y_{ijk}$ as being

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j. \tag{15}$$

This means that the total effect of variety $i$ and treatment $j$ on expected yield is considered as being the sum of the two individual effects $\alpha_i$ and $\beta_j$. For this reason, the effects are described as being *additive*. The model also means that the effect of variety $i$ on expected yield is considered the same no matter what treatment is used on it. For all treatments, the effect of variety $i$ is assumed to be $\alpha_i$ and the combined effect of variety $i$ and treatment $j$ over and above $\mu$ is taken to be $\alpha_i + \beta_j$.

Suppose hypothetical values of the $\mu$, the $\alpha_i$'s and the $\beta_j$'s are taken to be

$$\begin{array}{llll} \mu = 4, & \alpha_1 = 1 & \text{and} & \beta_1 = 4 \\ & \alpha_2 = 3 & & \beta_2 = 7. \\ & \alpha_3 = 2 & & \end{array} \tag{16}$$

The values of the $\mu$, the $\alpha_i$'s and the $\beta_j$'s are not observed values. They are introduced for purposes of illustration. In fact, these elements can never be observed. In practice they are never known because they are population values that can only be estimated from observed data. However, for purposes of illustrating, certain aspects of linear models it is instructive to give numerical values to these elements and portray the results graphically. For example, with the assumed values of (16),

$$E(y_{11k}) = \mu + \alpha_1 + \beta_1 = 4 + 1 + 4 = 9. \tag{17}$$

This is not an observed value of $E(y_{11k})$ or of $y_{11k}$ itself. It is an assumed value of $E(y_{11k})$ based on the assumed values of the parameters given in (16).

First note that (15) for a given $i$ and $j$ is the same for all $k$. Since the subscript $k$ is merely the identifier of the individual observations in the $(i, j)$ subclass, (15) means that the expected value of every observation in that subclass is the same. Thus, by (17), the expected value of every observation in the (1, 1) cell is, in our hypothetical example 9. That means that for all $k = 1, 2, \ldots, n_{11}$, $E(y_{11k}) = 9$. With this interpretation, the expected values for the other subclasses derived from (16) are those shown in Table 4.4 and plotted in Figure 4.1.

In Figure 4.1, note that the "variable" of the abscissa, the variety number, is not a continuous variable. Lines joining the values of $E(y_{ijk})$ do not indicate a continuous change in $E(y_{ijk})$ from one variety to the next. The lines are shown merely to emphasize the trend of the change. They are used in a similar way in Figures 4.2, 4.3, and 4.4. Furthermore, the ordinates plotted in these figures are values of $E(y_{ijk})$ and not of

**TABLE 4.4 Expected Values in a No-Interaction Model Equations (16) Substituted in (15) (See Figures 4.1 and 4.3)**

| | Treatment | |
|---|---|---|
| Variety | 1 | 2 |
| 1 | $E(y_{11k}) = 4 + 1 + 4 = 9$ | $E(y_{12k}) = 4 + 1 + 7 = 12$ |
| 2 | $E(y_{21k}) = 4 + 4 + 4 = 11$ | $E(y_{22k}) = 4 + 3 + 7 = 14$ |
| 3 | $E(y_{31k}) = 4 + 2 + 4 = 10$ | $E(y_{32k}) = 4 + 2 + 7 = 13$ |

actual observations $y_{ijk}$. With this in mind, it is clear from Figure 4.1 that in the hypothetical example of the model given in (15), the effect of variety is the same regardless of treatment. For *both* treatments, variety 2 has an expected yield two units larger than does variety 1. For both treatments, variety 3 is one unit lower than variety 2.

**(*ii*) *Interactions.*** In some other hypothetical example, suppose that the plots of expected yields are those shown in Figure 4.2. Observe that in Figure 4.1, the lines for the two treatments were parallel and that in Figure 4.2, they are not. This indicates that the effect of variety is different for different treatments. For treatment 1, variety 2 has an expected yield, that is, three units *larger* than that of variety 1 for the same treatment. However, for treatment 2, the expected yield of variety 2 is four units *smaller* than that of variety 1. Thus, in this second hypothetical example, the

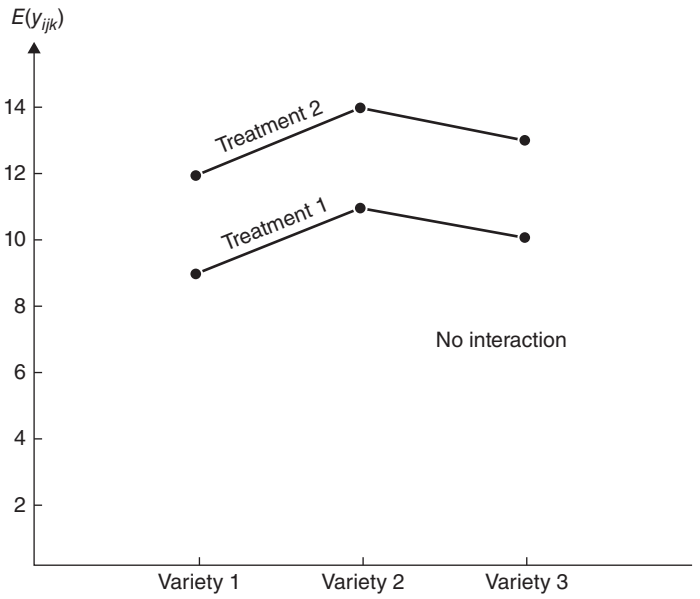



**FIGURE 4.1** Expected Values of Table 4.4

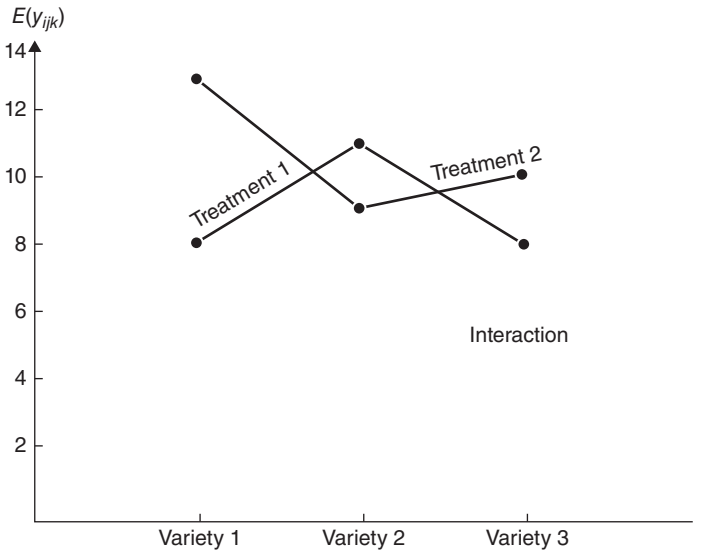



**FIGURE 4.2** Expected Values for an Interaction Model (See Table 4.5)

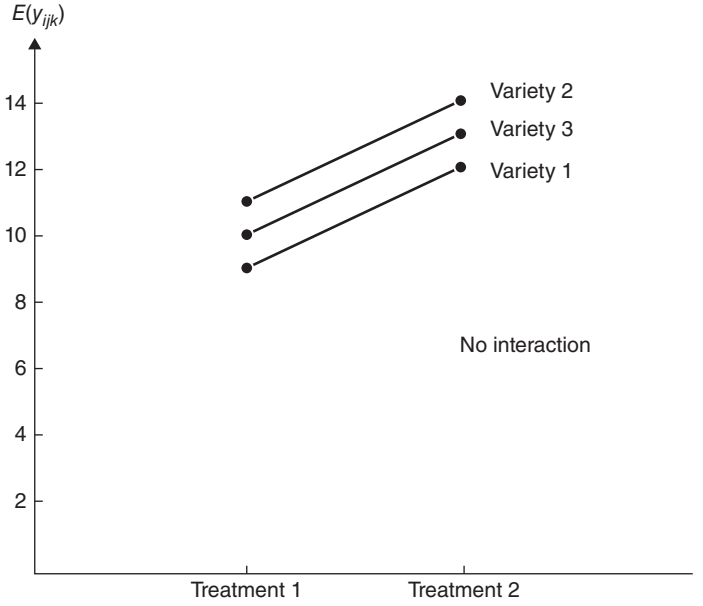



**FIGURE 4.3** Expected Values of Table 4.4 (See Also Figure 4.1)

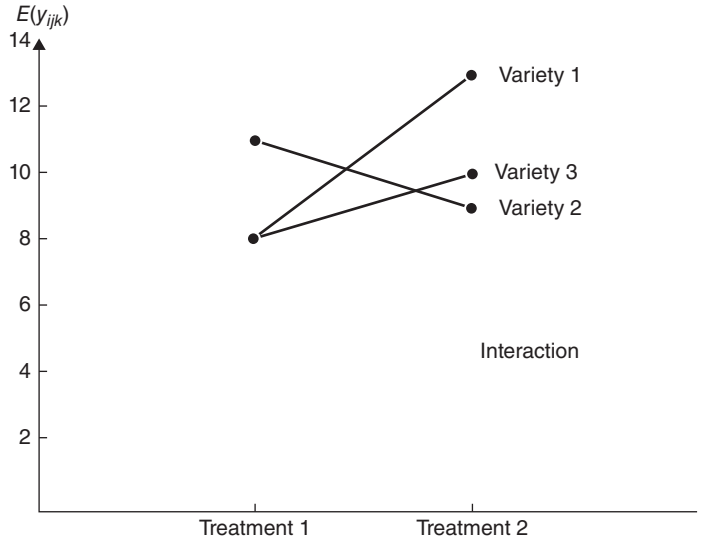



**FIGURE 4.4** Expected Values for an Interaction Model (See Table 4.5 and Figure 4.2)

varieties are acting differently for the different treatments. We say that the varieties are "interacting" with treatments. The extent to which they are not acting in the same manner for each treatment is termed an "interaction."

We can look at this in another way. In Figure 4.1, the difference between treatments is the same for each variety. It does not change from variety to variety. It is constant over all varieties. We can see this from the parallelism of the lines in Figure 4.1. On the other hand, the lack of parallelism in Figure 4.2 indicates that the differences between treatments are in fact different from variety to variety. Thus, the difference "treatment 1 minus treatment 2" is –5, +2, and –2 for the three varieties, respectively. However, in Figure 4.1, the same difference is –3 for every variety. This difference is well illustrated by plotting them in Figures 4.3 and 4.4.

The parallel lines in Figure 4.3 (corresponding to those in Figure 4.1) illustrate for the first hypothetical example (Table 4.4), the uniform difference between treatments of all varieties. However, the non-parallel lines in Figure 4.4 illustrate, for the second hypothetical example, the lack of uniformity in the differences between treatments over all varieties.

It is evident, from this discussion, that in Figures 4.1 and 4.3 (Table 4.4), the effect of variety on expected yield is the same for all treatments and that the effect of treatments is the same for all varieties. This also follows from the form of equation (15) used as a basis for Table 4.4 and Figures 4.1 and 4.3. However, in Figures 4.2 and 4.4, the effect of treatment is not the same for all varieties, and the effect of varieties is not the same for all treatments. This indicates that there are some additional effects accounting for the way those treatments and varieties are interacting. These additional effects are called interaction effects. They represent the manner in which one level of each main effect (variety) interacts with each level of the other main effect (treatment).

We take these effects into account in the equation of the model by adding another term. Thus, if the interaction effect between the $i$th level of the $\alpha$-effect and the $j$th level of the $\beta$-effect is $\gamma_{ij}$, the equation of the model is

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{18}$$

or equivalently

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \tag{19}$$

All other elements have the same meaning as before.

The second hypothetical example (plotted in Figures 4.2 and 4.4) is based on the same hypothetical values for $\mu$, the $\alpha$'s and the $\beta$'s given in (16) together with the following hypothetical values for the interaction effects $\gamma_{ij}$.

$$\begin{array}{ll} \gamma_{11} = 1 & \gamma_{21} = 1 \\ \gamma_{12} = 0 & \gamma_{22} = -5 \\ \gamma_{13} = -2 & \gamma_{31} = -3. \end{array} \tag{20}$$

In this way, the expected values derived from (18) are those shown in Table 4.5 and plotted in Figures 4.2 and 4.4.

Notice that this description of interactions is entirely in terms of expected yields, that is, in terms of models having interactions in them. Such models may be used whenever we think that the data being dealt with behave in the manner illustrated. However, the simple numbers used in the example do not refer to data. They merely exemplify a model.

Note that whenever $n_{ij} = 1$ for all cells, the model with interaction (19) becomes indistinguishable from the model without interaction, (13). The $\gamma_{ij}$ and $e_{ijk}$ terms of (19) get combined, $\gamma_{ij} + e_{ijk} = \varepsilon_{ij}$, say, and so (19) becomes

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

equivalent to (13). This means that when $n_{ij} = 1$, we generally can study only the no-interaction model and not the interaction model.

**TABLE 4.5 Expected Values of an Interaction Model. Equations (16) and (20) Substituted in (18). (See Figures 4.2 and 4.4)**

| | Treatment | |
|---|---|---|
| Variety | 1 | 2 |
| 1 | $E(y_{11k}) = 4 + 1 + 4 - 1 = 8$ | $E(y_{12k}) = 4 + 1 + 7 + 1 = 13$ |
| 2 | $E(y_{21k}) = 4 + 3 + 4 + 0 = 11$ | $E(y_{22k}) = 4 + 3 + 7 - 5 = 9$ |
| 3 | $E(y_{31k}) = 4 + 2 + 4 - 2 = 8$ | $E(y_{32k}) = 4 + 2 + 7 - 3 = 10$ |

For a specific kind of interaction when $n_{ij} = 1$, Tukey (1949) developed a test for non-additivity. For a discussion of this test, see, for example, Rao (1973), pp. 249–251.

The interactions that have been discussed may be generalized. The $\gamma_{ij}$ is an interaction between two factors and is known as a *first-order interaction*. An interaction between three factors is called a *second-order interaction*. Third, fourth, and higher order interactions follow in like manner. The interpretation becomes more difficult as the order of interactions increase. For example, a third-order interaction (which involves the interaction between four main effects) can be interpreted as the interaction between a main effect and a second-order interaction or as the interaction between two first-order interactions.

**Notation.** A frequently used notation that helps clarify the interpretation of interactions is based on using the symbol $(\alpha\beta)_{ij}$ in place of $\gamma_{ij}$. This indicates that $(\alpha\beta)_{ij}$ is the interaction effect between the $i$th level of the $\alpha$-factor and the $j$th level of the $\beta$-factor. The symbol $(\alpha\beta)_{ij}$ in *no way* indicates the product of any $\alpha$ with any $\beta$ even if it is written without parenthesis as $\alpha\beta_{ij}$. It is a combined symbol that indicates more clearly than $\gamma_{ij}$ that it represents an interaction between a $\alpha$-factor and a $\beta$-factor. By this means a higher order interaction $(\alpha\beta\gamma\delta)_{hijk}$, for example, may be readily interpreted. It may be thought of as the interaction between $\alpha_h$ and $(\beta\gamma\delta)_{ijk}$. It could also be interpreted as the interaction between $(\alpha\beta)_{hi}$ and $(\gamma\delta)_{jk}$. There are many other interpretations of this interaction. This notation also clarifies that the writing of a model

$$y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \tau_{ij} + \rho_{ik} + \theta_{jk} + \varphi_{ijk} + e_{ijkm}$$

is not as easily understood as when the model is written as

$$y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkm}.$$

Finally, even when a model has interactions, its order is still described by the number of main effects it has. Thus, (18) is an equation for a two-way classification model, just as is (13). However, (18) includes interactions and (13) does not.

### e. Nested and Crossed Classifications

In the example of Table 4.3, every treatment is applied to every variety. Even though there are no observations on the use of treatment 2 with variety 3, this combination would be feasible if there were data available. This kind of situation is called a *crossed classification*. In a crossed classification, every level of every factor can be used with every level of every other factor. In this way, the factors "cross" each other. Their "intersections" are the subclasses or the cells of the situation, wherein data arise. Absence of data from a cell does not imply non-existence of that cell. It only implies that that cell has no data. The total number of cells in a crossed classification is the

product of the number of product of the number of levels of the various factors. For example, if we had two treatments and three varieties, there would be six cells. Not all of the cells need to have observations in them. If, say, $s$ cells have observations in them, then the total number of observations would be the sum of the number of observations in the $s$ cells.

We have already had an example of a crossed classification. We now give an example to introduce nested classifications.

**Example 3 A Nested Classification** Suppose at a university, a student survey is carried out to ascertain the reaction to instructor's use of the computer in teaching different courses. Suppose that all freshmen have to take English, Geology, or Chemistry their first semester and one other of these courses their second semester. All three first semester courses are large and are divided into sections with different instructors. The sections need not have the same number of students. In the survey, the response provided by each student measures his opinion of his instructors use of the computer on a scale of 1 to 10. The questions of interest are

1. Do instructors differ in their use of the computer?
2. Is the use of the computer affected by the subject matter being taught?

A possible model for the situation would include a general mean $\mu$ and main effects $\alpha_1, \alpha_2$, and $\alpha_3$ for the three types of courses. It would also include terms for the sections of each course. Assume that there are 10 sections for each course. We try to use a model of the form $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$ for $i = 1, 2, 3, j = 1, 2, \ldots, 10$, and $k = 1, 2, \ldots, n_{ij}$. The number $n_{ij}$ represents the number of students in section $j$ of course $i$. Consider $\beta_j$; it represents the effect of section $j$.

For $j = 1$, say, it would be the effect of section 1 of the English course, the Geology course, and the Chemistry course. This is meaningless, because these three sections composed of different groups of students, have nothing in common other than they are all numbered as 1 in their respective courses. However, assuming students in all courses have been allocated randomly to their sections, this numbering is purely for identification purposes. It indicates nothing in common about the three sections that are numbered 1. Neither is there anything in common about the three sections that are numbered 5, or 6, or any other number. They are not like the treatments in the agricultural example, where treatment 1 on variety 1 was the same as treatment 1 on variety 2 and on variety 3. The sections are not related in this way. They are identities in their own courses. Thus, we refer to them as sections within courses, and describe them as being *nested* within courses. Thus, sections are a nested factor, or a *nested classification*, sometimes also referred to as a *hierarchical* classification.

The difference between a crossed classification and a nested classification is exemplified in Table 4.6, in terms of the variety and treatment example described earlier, and the sections-within-courses example just discussed.

**TABLE 4.6 Schematic Representation of a Crossed Classification and a Nested Classification**

| A Crossed Classification | | | A Nested Classification | | |
|---|---|---|---|---|---|
| | Treatment | | Course | | |
| Variety | 1 | 2 | English | Geology | Chemistry |
| 1 | | | Section 1 | Section 1 | Section 1 |
| 2 | | | Section 2 | Section 2 | Section 2 |
| 3 | | | Section 3 | | Section 3 |
| | | | | | Section 4 |

In the crossed classification, variety 1 is used in combination with both treatment 1 and treatment 2, and it is the same variety on both occasions. In the nested classification, section 1 of English is in no way related to section 1 of Geology. The only thing in common is the number 1, which is purely an identifier. In the crossed classification, every level of the one factor is used in combination with every level of the other factor. However, in the nested classification the levels of the nested factors (sections) are unrelated to one another and are nested within a level of the other factor. Further as illustrated in Table 4.6, there may be different numbers of levels of the nested factor within each level of the other factor (different numbers of sections in the different courses).

The equation of the model accounts for the nesting of sections within courses by giving to the effect $\beta_j$ for the $j$th section the subscript $i$, for course, so that $\beta_{ij}$ is then the effect for the $j$th section nested within the $i$th course. This signifies that the $j$th section cannot be defined alone but only in context of which course it belongs to. Thus, the model is

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk} \tag{21}$$

where $y_{ijk}$ is the opinion of student $k$ in the $j$th section of course $i$. The limits of $k$ are $k = 1, 2, \ldots, n_{ij}$ where there are $n_{ij}$ students in the $j$th section of the $i$th course, and $j = 1, 2, \ldots, b_i$ where there are $b_i$ students in course $i$, and $i = 1, 2, 3$. Table 4.7 summarizes a situation of a total of 263 students in 3 sections in English, 2 in Geology, and 4 in Chemistry.

**TABLE 4.7 A Nested Classification**

| English ($i = 1$)<br>3 Sections, $b_1 = 3$ | Geology ($i = 2$)<br>2 Sections, $b_2 = 2$ | Chemistry ($i = 3$)<br>4 Sections, $b_3 = 4$ |
|---|---|---|
| $n_{11} = 28$ | $n_{21} = 31$ | $n_{31} = 27$ |
| $n_{12} = 27$ | $n_{22} = 29$ | $n_{32} = 32$ |
| $n_{13} = 30$ | | $n_{33} = 29$ |
| | | $n_{34} = 30$ |

The situation illustrated in Table 4.7 is described as a two-way nested classification: sections within courses. Now, consider asking a student his opinion on two different occasions, say right after the midterm examinations and after he/she completes the course. If $y_{ijkh}$ is the $h$th reply ($h = 1$ or 2) of the $k$th student in section $j$ of course $i$, a suitable model might be

$$y_{ijkh} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + e_{ijkh}. \tag{22}$$

Now we have not only sections listed within courses but also students nested within sections. Students are nested within sections for exactly the same reason that sections are nested within courses. In general, there is no limit to the degree of nesting that can be handled. The extent of its use depends entirely on the data and the environment from which they came.

**Notation.** The meaning of the term $\gamma_{ij}$ in (19) might, at first sight, be confused with the meaning of $\beta_{ij}$ in (21), although the context of each does make their meaning clear. By the presence of $\alpha_i$ and $\beta_j$ in (19), $\gamma_{ij}$ is clearly an interaction effect. By the lack of a term with just a subscript $j$ from (21), it is clear that $\beta_{ij}$ is a term for a nested factor. However, additional clarity can be brought to the situation by using the $(\alpha\beta)_{ij}$-notation for interactions (as already described), for then $\beta_{ij}$ is clearly a main effect. Similar clarity is gained by using $\beta_{(i)j}$ instead of $\beta_{ij}$ for a nested effect. This makes it clear that $\beta_{(i)j}$ is not an interaction effect like $\gamma_{ij}$. Either or both of these notations can be used to insure against confusion.

Interaction effects are effects peculiar to specific combinations of the factors involved. Thus, $(\alpha\beta)_{ij}$ is the interaction effect peculiar to the combination of the $i$th level of the $\alpha$-factor and the $j$th level of the $\beta$-factor. Interactions between a factor and one nested within it cannot, therefore exist. This is so because, for example, when sections are nested within courses they are defined only within that context. There is no such thing as a section factor in which exactly the same level occurs in combination with the levels of the course factor. For example, section 1 as defined for English never occurs in combination with Chemistry that has its own section 1. As a result, there is no such thing as an interaction between courses and sections within courses. The notation of (21) makes this quite clear. The interaction between $\alpha_i$ and $\beta_{ij}$ would be $(\alpha\beta)_{ij}$ which cannot be identified separately from $\beta_{ij}$. Therefore, there is no interaction.

Nested and crossed classifications are by no means mutually exclusive. Both can occur in the same model. For example, in using (22) as the model for the repeated surveying of the students, we are ignoring the fact that the two surveys (assumed to be conducted from the same questionnaire) will have to be made at different times. If the time element were to be included a suitable model would be

$$y_{ijkh} = \mu + \delta_h + \alpha_i + (\alpha\delta)_{ik} + \beta_{ij} + \gamma_{ijk} + e_{ijkh} \tag{23}$$

where all terms are the same as those previously with the addition of the term $\delta_h$, the effect of time $h$. The $\delta$-factor (time) and the $\alpha$-factor (courses) are crossed factors because each level of one occurs with every level of the other. As before, the $\beta$-factor (sections within courses) and the $\gamma$-factor (students within sections) are nested factors. Interactions could be included in a model for this situation too. The model

$$y_{ijkh} = \mu + \delta_h + \alpha_i + (\alpha\delta)_{ih} + \beta_{ij} + \gamma_{ijk} + e_{ijkh} \tag{24}$$

includes a term for the interaction between time and course.

We can see that the variations that may be rung on the same theme are very numerous. Just what goes into a model depends, of course, on the nature of the data to be analyzed, the things of interest to the researcher and the assumptions that he is prepared to make. For example, if time is to be ignored, either by assumption or because it is known to be of no importance, then (22) would be an acceptable model. Even so, it might be questioned whether or not we truly know that something is of no importance, and in this light maybe model (23) or (24) should be used. On the other hand, if the second student survey had been carried out following a major modification to the computer system designed to improve its efficiency and attractiveness to instructors, there is no question that (22) would be an unsuitable model compared to (23) and (24). This is because $\delta_1$ and $\delta_2$ would then represent the effects of the two computer systems modified and unmodified.

On all occasions, the environment in which the data were gathered determines the model.

In conclusion, it is to be emphasized that all these kinds of models can be written as $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ just as they were in equations (4) or (8). For all of them $\mathbf{X}$ will have 0's and 1's for its elements and not be of full column rank. However, for all these models, the estimation procedure of Chapter 3 can be used to derive the normal equations $\mathbf{X'X}\hat{\mathbf{b}} = \mathbf{X'y}$. In these, $\mathbf{X'X}$ does not have full rank. Nevertheless, the equations can be solved using a generalized inverse of $\mathbf{X'X}$. This and its consequences are discussed in detail in Chapter 5. As a prelude, we consider a numerical example to illustrate some points involved.

## 4. THE NORMAL EQUATIONS

The equation of the general linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is identical to that used for regression analysis in Chapter 3. There the normal equations for estimating $\mathbf{b}$ were written as $\mathbf{X'X}\hat{\mathbf{b}} = \mathbf{X'y}$, where $\hat{\mathbf{b}}$ was the estimator of $\mathbf{b}$. The same kind of normal equations can be used here. However, we will write them as $\mathbf{X'Xb}^\circ = \mathbf{X'y}$. This is done because there is no single solution for $\mathbf{b}^\circ$. The matrix $\mathbf{X'X}$ is singular and so the normal equations have infinitely many solutions. None of these solutions is an estimator of $\mathbf{b}$ in the sense that $\hat{\mathbf{b}}$ is in regression analysis, and so we introduce the symbol $\mathbf{b}^\circ$. It represents a solution to the normal equations but is not an estimator of $\mathbf{b}$. This point is emphasized repeatedly in Chapter 5 and is illustrated by the introduction therein.

Suppose the values of the six observations in Table 4.2 are, in some appropriate unit of measurement.

$$\mathbf{y}' = [\,16 \quad 10 \quad 19 \quad 11 \quad 13 \quad 27\,].$$

Comparable to $\mathbf{b}$ in (5) we now use

$$\mathbf{b}' = [\,\mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3\,],$$

where $\mu$ is a general mean and the $\alpha$'s are effects due to educational status. Then with $\mathbf{X}$ of (5) the normal equations are

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \end{bmatrix} = \begin{bmatrix} 96 \\ 45 \\ 24 \\ 27 \end{bmatrix}. \tag{25}$$

These are equivalent to

$$\begin{aligned} 6\mu^\circ + 3\alpha_1^\circ + 2\alpha_2^\circ + \alpha_3^\circ &= 96 \\ 3\mu^\circ + 3\alpha_1^\circ &= 45 \\ 2\mu^\circ + 2\alpha_2^\circ &= 24 \\ \mu^\circ + \alpha_3^\circ &= 27 \end{aligned}$$

The derivation of equations such as these will be discussed in Chapter 5. All we note here is that the sum of the last three equals the first and hence they have infinitely many solutions. Four of these are shown in Table 4.8.

The differences between the same elements of the four solutions shown in Table 4.8 make it crystal clear why no solution $\mathbf{b}^\circ$ can be considered as an estimator of $\mathbf{b}$. For this reason, $\mathbf{b}^\circ$ is always referred to as a solution to normal equations and never as an estimator. The notation $\mathbf{b}^\circ$ emphasizes this, distinguishing it from $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_c$ of equations (21) and (118) of Chapter 3.

An investigator having data to be analyzed will clearly have no use for $\mathbf{b}^\circ$ as it stands, whatever its numerical value is. However, what about linear functions of

**TABLE 4.8 Four Solutions $\mathbf{b}_1^\circ, \mathbf{b}_2^\circ, \mathbf{b}_3^\circ$, and $\mathbf{b}_4^\circ$ to Equations (25)**

| | Solution | | | |
|---|---|---|---|---|
| Element of Solution | $\mathbf{b}_1^\circ$ | $\mathbf{b}_2^\circ$ | $\mathbf{b}_3^\circ$ | $\mathbf{b}_4^\circ$ |
| $\mu^\circ$ | **16** | **14** | **27** | **−2982** |
| $\alpha_1^\circ$ | **−1** | **1** | **−12** | **2997** |
| $\alpha_2^\circ$ | **−4** | **−2** | **−15** | **2994** |
| $\alpha_3^\circ$ | **11** | **13** | **0** | **3009** |

**TABLE 4.9 Values of $\frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ)$ and $(\mu^\circ + \alpha_1^\circ + \alpha_1^\circ + \alpha_1^\circ)/3$**

| | Solution (See Table 4.8) | | | |
|---|---|---|---|---|
| Linear function | $b_1^\circ$ | $b_2^\circ$ | $b_3^\circ$ | $b_4^\circ$ |
| $\frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ)$ | 3.5 | 5.5 | −7.5 | 3001.5 |
| $(\mu^\circ + \alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ)/3$ | 7.333 | 8.667 | 0 | 2006 |

the elements of $\mathbf{b}^\circ$? Suppose, for example, there is interest in estimating the mean effect on investment of high school and of college education? Recall from Table 4.2 that corresponding to $\alpha_1, \alpha_2$, and $\alpha_3$ in the model are the three levels of educational status, high school incomplete, high school graduate, and college graduate, and the $y$-variable is investment in consumer durables. This suggests the following question. Even if $\mathbf{b}^\circ$ is of no use in itself what does it do for values such as $\frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ)$ or for $(\mu^\circ + \alpha_1^\circ + \alpha_1^\circ + \alpha_1^\circ)/3$? The answer is seen in Table 4.9. Exactly as with the elements of $\mathbf{b}^\circ$ itself in Table 4.8, the values in Table 4.9 vary greatly from solution to solution.

Fortunately, this is not the case for all linear functions. There are linear functions of the solutions to normal equations that an investigator might be interested in that have the same numerical value regardless of which set of solutions is used. Examples are given in Table 4.10.

Notice that each of the linear functions is invariant to the solution $\mathbf{b}^\circ$ that is used. Since this is so for all of the infinitely many solutions $\mathbf{b}^\circ$, these expressions are of value to the investigator. Moreover, by their nature, these expressions are often those of specific interest to the investigator because they can be described as follows:

$\alpha_1^\circ - \alpha_2^\circ$ : estimator of difference between effects of two levels.
$\mu^\circ + \alpha_1^\circ$ : estimator of general mean plus mean effect of a level.
$\mu^\circ + \frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ)$: estimator of general mean plus mean effect of two levels.
$\frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ) - \alpha_1^\circ$: estimator of superiority of main effect of two levels over the effect of another level.

**TABLE 4.10 Estimates of Four Estimable Functions**

| | Solution (See Table 4.8) | | | |
|---|---|---|---|---|
| Linear function | $b_1^\circ$ | $b_2^\circ$ | $b_3^\circ$ | $b_4^\circ$ |
| $\alpha_1^\circ - \alpha_2^\circ$ | 3 | 3 | 3 | 3 |
| $\mu^\circ + \alpha_1^\circ$ | 15 | 15 | 15 | 15 |
| $\mu^\circ + \frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ)$ | 19.5 | 19.5 | 19.5 | 19.5 |
| $\frac{1}{2}(\alpha_2^\circ + \alpha_3^\circ) - \alpha_1^\circ$ | 4.5 | 4.5 | 4.5 | 4.5 |

Of course, these are only four of the many such linear functions of elements of $\mathbf{b}^\circ$ having the property demonstrated in Table 4.10. Other similar linear functions include $\alpha_3^\circ - \alpha_1^\circ, \mu^\circ + \alpha_2^\circ, \mu^\circ + \frac{1}{2}(\alpha_1^\circ + \alpha_2^\circ)$ and infinitely many others. Functions such as these are known as *estimators of estimable functions*. They all have the property that they are invariant to whatever solutions are obtained to the normal equations. Because of this invariance property, they are the only functions that can be of interest, so far as estimation of the parameters of a linear model is concerned. Distinguishing this class of functions from functions such as those illustrated in Table 4.9 that do not have the invariance property is important as is deriving their other properties. This will be taken up in Chapter 5.

## 5. EXERCISES

The following statement and Table 4.11 apply to Exercises 1–10.

Suppose an oil company gets its crude oil from four different sources, refines it in three different refineries using the same two processes in each refinery. In one part of the refining process, a measurement of efficiency is taken as a percentage and recorded as an integer between 0 and 100. Table 4.11 shows the available measurement of efficiency for different samples of oil.

**1** Write the equation of the linear model for each of the following cases

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

for considering the effect of refinery and process on efficiency giving the explicit forms of **y, X, b,** and **e.**

**(a)** The eight observations on Texas.

**(b)** The six observations on Oklahoma.

**(c)** The five observations for the Gulf of Mexico.

**(d)** The six observations for the Iran data.

**TABLE 4.11 Results of Efficiency Tests**

| | | Source | | | |
|---|---|---|---|---|---|
| Refinery | Process | Texas | Oklahoma | Gulf of Mexico | Iran |
| Galveston | 1 | 31, 33, 44, 36 | 38 | 26 | — |
| | 2 | 37, 59 | 42 | — | — |
| Newark | 1 | — | — | 42 | 34, 42, 28 |
| | 2 | 39 | 36 | 32, 38 | — |
| Savannah | 1 | 42 | 36 | — | 22 |
| | 2 | — | 42, 46 | 26 | 37, 43 |

**2** Repeat Exercise 1 including interactions between refinery and process.

**3** For Exercise 1a,

**(a)** Write the normal equations.

**(b)** Show that two solutions to these normal equations are

$$\mathbf{b}_1^\circ = \left[\begin{array}{cccccc} \frac{246}{11} & \frac{93}{11} & \frac{-6}{11} & \frac{159}{11} & \frac{57}{11} & \frac{189}{11} \end{array}\right]' \text{ and } \mathbf{b}_2^\circ = \left[\begin{array}{cccccc} 0 & 0 & -9 & 6 & 36 & 48 \end{array}\right]'.$$

**(c)** Show that for these two solutions that $\alpha_1^\circ - \alpha_2^\circ, \alpha_1^\circ - \alpha_3^\circ, \alpha_2^\circ - \alpha_3^\circ, \beta_1^\circ - \beta_2^\circ$ are the same but that $\mu^\circ + \alpha_i^\circ, i = 1, 2, 3, \mu^\circ + \beta_j^\circ, j = 1, 2$ are not.

**4** Consider a linear model with

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right], \mathbf{Y} = \left[\begin{array}{c} 31 \\ 33 \\ 44 \\ 36 \\ 38 \\ 26 \end{array}\right], \mathbf{b} = \left[\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{array}\right]$$

**(a)** Write the normal equations.

**(b)** Using the methods of Chapter 1, find two generalized inverses of $\mathbf{X}'\mathbf{X}$.

**(c)** Use your answer to (b) to find two estimates $\mathbf{b}^\circ$.

**(d)** Find two linear combinations of the elements of $\mathbf{b}^\circ$ that are the same for both estimates and two that are different.

**5** Write down the linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ giving the explicit forms of **X, b,** and **e** for each of the following refineries.

**(a)** Galveston (nine observations)

**(b)** Newark (eight observations)

**(c)** Savannah (eight observations)

**6** Write down the linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ giving the explicit forms of **X, b,** and **e** for each of the two processes. Let the $\alpha_i$'s be the sources and the $\beta_j$'s be the refineries.

**7** **(a)** For all 25 observations in Table 4.11, write down the equations of the linear model for considering the effect of source, refinery, and process on efficiency. Do not include interactions.

**(b)** Write down the normal equations.

**8** For the Texas data.

**(a)** Write the linear model assuming that process is nested within refinery.

**(b)** What are the normal equations?

**9** Use the data for Texas–Galveston, Texas–Savannah, Oklahoma–Galveston, and Oklahoma–Savannah omitting process. Just consider the data from two different processes as replications. Write down the linear models in matrix form for the following situations.

**(a)** The two-way model without interaction.

**(b)** The two-way model with interaction.

**(c)** The model with refineries nested within the source.

**10** For Exercise 9a

**(a)** Obtain the normal equations

**(b)** Find two solutions using generalized inverses.

**(c)** Give two linear combinations where results are the same and two where they are different.

# 5

# MODELS NOT OF FULL RANK

Chapter 3 discussed regression analysis for a model having equation $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $\mathbf{X}$ has full-column rank. Chapter 4 illustrated how the same equation can apply to linear models when $\mathbf{X}$ may not have full-column rank. We shall now consider estimation and hypothesis testing for the non-full-rank case. We will follow the same sequence of development as in Chapter 3. In Chapter 4, we gave an informal demonstration of estimable functions. In this chapter, we will formally define them and study their important properties.

## 1. THE NORMAL EQUATIONS

As before, we deal with the model,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}.$$

Again, $\mathbf{y}$ is an $N \times 1$ vector of observations $y_i$, $\mathbf{b}$ is a $p \times 1$ vector of parameters, $\mathbf{X}$ is an $n \times p$ matrix of known values (in most cases 0's and 1's) and $\mathbf{e}$ is a vector of random error terms. As before, $\mathbf{e}$ is defined as

$$\mathbf{e} = \mathbf{y} - E(\mathbf{y}).$$

Then $E(\mathbf{e}) = \mathbf{0}$ and $E(\mathbf{y}) = \mathbf{Xb}$. Every element in $\mathbf{e}$ is assumed to have variance $\sigma^2$ and zero covariance with every other element. More formally, we have

$$\text{var}(\mathbf{e}) = E(\mathbf{ee}') = \sigma^2 \mathbf{I}_N.$$

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

Thus,

$$\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I}) \quad \text{and} \quad \mathbf{y} \sim (\mathbf{Xb}, \sigma^2\mathbf{I}).$$

Normality will be introduced later when we consider hypothesis testing and confidence intervals.

### a. The Normal Equations

As was done in Chapter 3, the normal equations corresponding to the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is derived using least squares. As before, when $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$, the normal equations turn out to be

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}. \tag{1}$$

We shall discuss the more general case where $\text{var}(\mathbf{e}) = \mathbf{V}$, whether $\mathbf{V}$ be singular or non-singular in Section 8.

Before solving equations (1), we look at their form, initially in terms of an illustrative example.

**Example 1 Finding the Normal Equations** Deoxyribonucleic acid (DNA) is the hereditary material found in most organisms. A genome is an organism's complete set of DNA, including all of its genes. Each genome contains all the information needed to build and maintain that organism.

Macdonald (2015) presents an example with genome size measured in pictograms (trillionths of a gram) of DNA per haploid cell in several large groups of crustaceans. The data are taken from Gregory (2015). For purposes of illustration, we shall consider six points for three kinds of crustaceans. We shall also use these data for subsequent examples to give numerical illustrations of the computations. Searle (1971) presents similar examples for data on rubber tree plants, taken from Federer (1955).

For the entries in Table 5.1, let $y_{ij}$ denote the DNA content of the $j$th crustacean of the $i$th type, $i$ taking values 1, 2, and 3 for amphipods, barnacles, and branchiopods, respectively, and $j = 1, 2, \ldots, n_i$, where $n_i$ is the number of observations of the $i$th

**TABLE 5.1 Amount of DNA**

| | Type of Crustacean | | |
|---|---|---|---|
| | Amphipods | Barnacles | Branchiopods |
| | 27.00 | 0.67 | 0.19 |
| | 50.91 | 0.90 | |
| | 64.62 | | |
| | – | – | – |
| Total | 142.53 | 1.57 | 0.19 |

type. The objective is to estimate the effect of the type of crustacean DNA content. To do this, we assume that the observation $y_{ij}$ is the sum of three parts

$$y_{ij} = \mu + \alpha_i + e_{ij},$$

where $\mu$ represents the population mean of the DNA content of the crustaceans, $\alpha_i$ is the effect of type $i$ DNA content, and $e_{ij}$ is a random error term peculiar to the observation $y_{ij}$.

To develop the normal equations, we write down the six observations in terms of the equation of the model

$$\begin{aligned}
27.00 = y_{11} &= \mu + \alpha_1 + e_{11} \\
50.91 = y_{12} &= \mu + \alpha_1 + e_{12} \\
64.62 = y_{13} &= \mu + \alpha_1 + e_{13} \\
0.67 = y_{21} &= \mu + \alpha_2 + e_{21} \\
0.90 = y_{22} &= \mu + \alpha_2 + e_{22} \\
0.19 = y_{31} &= \mu + \alpha_2 + e_{31}.
\end{aligned}$$

We may rewrite these equations in matrix form $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ as

$$\begin{bmatrix} 27.00 \\ 50.91 \\ 64.62 \\ 0.67 \\ 0.90 \\ 0.19 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \end{bmatrix} \tag{2}$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{X}$ is the matrix of 0's and 1's, $\mathbf{b}$ is the vector of parameters to be considered,

$$\mathbf{b}' = [\mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3],$$

and $\mathbf{e}$ is the vector of error terms. □

The vector $\mathbf{b}$ in $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is the vector of parameters. It is the vector of all of the elements of the model. In this case, the elements are $\mu, \alpha_1, \alpha_2$, and $\alpha_3$. This representation holds true in general for all linear models. For example, if data can be arranged in rows and columns according to two different classifications, the vector $\mathbf{b}$ will have as its elements the term $\mu$, the terms representing row effects, those representing column effects, and those representing interaction effects between rows and columns. For $r$ rows and $c$ columns, the vector $\mathbf{b}$ can have as many as $1 + r + c + rc = (1 + r)(1 + c)$ elements.

The matrix $\mathbf{X}$ in $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is called the *incidence matrix*, or sometimes the design matrix. This is because the location of the 0's and 1's throughout its elements

**TABLE 5.2 The Design Matrix X as a Two-Way Table**

| | Parameters of Model | | | |
|---|---|---|---|---|
| Observations | $\mu$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
| $y_{11}$ | 1 | 1 | 0 | 0 |
| $y_{12}$ | 1 | 1 | 0 | 0 |
| $y_{13}$ | 1 | 1 | 0 | 0 |
| $y_{21}$ | 1 | 0 | 1 | 0 |
| $y_{22}$ | 1 | 0 | 1 | 0 |
| $y_{31}$ | 1 | 0 | 0 | 1 |

represents the incidence of the terms of the model among the observations and hence of the classifications in which the observations lie. This is particularly evident if one writes $\mathbf{X}$ as a two-way table with the parameters as headings for the columns and the observations as labels for the rows, as illustrated in Table 5.2.

Consider the normal equations (1). They involve $\mathbf{X}'\mathbf{X}$ a square and symmetric matrix. Its elements are the inner products of the columns of $\mathbf{X}$ with each other. We have that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Since $\mathbf{X}$ does not have full-column rank, $\mathbf{X}'\mathbf{X}$ is not of full rank.

The normal equations also involve the vector $\mathbf{X}'\mathbf{y}$. Its elements are the inner products of the columns of $\mathbf{X}$ with the vector $\mathbf{y}$. Since the only non-zero elements of $\mathbf{X}$ are 1's, the elements of $\mathbf{X}'\mathbf{y}$ are certain sums of elements of $\mathbf{y}$. From (2),

$$\begin{aligned} \mathbf{X}'\mathbf{y} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} \\ &= \begin{bmatrix} y_{11} + y_{12} + y_{13} + y_{21} + y_{22} + y_{31} \\ y_{11} + y_{12} + y_{13} \\ y_{21} + y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix} = \begin{bmatrix} 144.29 \\ 152.53 \\ 1.57 \\ 0.19 \end{bmatrix}. \end{aligned} \tag{4}$$

In linear models, $\mathbf{X}'\mathbf{y}$ is frequently a vector of various subtotals of the $y$-observations.

As has already been mentioned in Chapter 4, when $\mathbf{X}'\mathbf{X}$ is not of full rank, as in (3), the normal equations (1) cannot be solved with one solitary solution $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as in Chapter 3. To emphasize this, we write the normal equations as

$$\mathbf{X}'\mathbf{X}\mathbf{b}^{\circ} = \mathbf{X}'\mathbf{y}. \tag{5}$$

We use the symbol $\mathbf{b}^{\circ}$ to distinguish the many solutions of (5) from the solitary solution that exists when $\mathbf{X}'\mathbf{X}$ is of full rank. We shall also use $\mathbf{b}^{\circ}$ to denote a solution $\mathbf{G}\mathbf{X}'\mathbf{y}$ to (5), where $\mathbf{G}$ is a generalized inverse $\mathbf{X}'\mathbf{X}$.

The normal equations of the crustacean example are, from (3) and (4),

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^{\circ} \\ \alpha_1^{\circ} \\ \alpha_2^{\circ} \\ \alpha_3^{\circ} \end{bmatrix} = \begin{bmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \\ y_{2\cdot} \\ y_{3\cdot} \end{bmatrix} = \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \end{bmatrix}. \tag{6}$$

By retaining the algebraic form of $\mathbf{X}'\mathbf{y}$ as well as its arithmetic form, we see that if $\mathbf{X}'\mathbf{X}$ is written in a two-way table, the row headings of the table will be the totals in $\mathbf{X}'\mathbf{y}$ and the column headings will be the parameters. Indeed, the elements of $\mathbf{X}'\mathbf{X}$ are the number of times that a parameter of a model occurs in a total. For example, $\mu$ occurs six times in $y_{\cdot\cdot}$ and $\alpha_1$ occurs three times. Likewise, $\alpha_2$ does not occur at all in $y_{1\cdot}$ and so on. Another way of looking at $\mathbf{X}'\mathbf{X}$ is that its elements are the coefficients of the parameters of the model in the expected values of the totals in $\mathbf{X}'\mathbf{y}$. In this sense, we might write the normal equations as $E(\widehat{\mathbf{X}'\mathbf{y}}) = \mathbf{X}'\mathbf{y}$ replacing $\mathbf{b}$ implicit in the left-hand side by $\mathbf{b}^{\circ}$. However, the easiest way of deriving $\mathbf{X}'\mathbf{y}$ and $\mathbf{X}'\mathbf{X}$ other than carrying out the matrix products explicitly is to form $\mathbf{X}'\mathbf{y}$ as the vector of all class and sub-class totals of the observations (including the grand total) and to form $\mathbf{X}'\mathbf{X}$ as the number of times that each parameter arises in each total that occurs in $\mathbf{X}'\mathbf{y}$.

### b. Solutions to the Normal Equations

Since $\mathbf{X}$ does not have full-column rank, $\mathbf{X}'\mathbf{X}$ has no inverse and the normal equations (5) have no unique solution. They have many solutions. To get any one of them, we find any generalized inverse $\mathbf{G}$ of $\mathbf{X}'\mathbf{X}$ and write the corresponding solution as

$$\mathbf{b}^{\circ} = \mathbf{G}\mathbf{X}'\mathbf{y}. \tag{7}$$

The ability to do this is a consequence of Theorem 11 in Chapter 1. We will use the results of Chapter 1 repeatedly in what follows, especially those of Section 5 of Chapter 1.

The notation $\mathbf{b}^{\circ}$ in equation (7) for a solution to the normal equation (5) emphasizes that what is derived in solving (5) is *only* a solution to the equation and *not* an estimator of $\mathbf{b}$. This point cannot be over-emphasized. In a general discussion of linear models

that are not of full rank, it is important to realize that what is obtained as a solution of the normal equations is just that, a solution and *nothing more*. It is misleading and, in most cases, quite wrong for $\mathbf{b}^\circ$ to be termed an estimator, particularly an estimator of $\mathbf{b}$. It is true that $\mathbf{b}^\circ$ is, as shall be shown, an estimator of something, but not of $\mathbf{b}$. Indeed the expression it estimates depends entirely upon which generalized inverse of $\mathbf{X}'\mathbf{X}$ is used in estimating $\mathbf{b}^\circ$. For this reason, $\mathbf{b}^\circ$ is always referred to as a solution and not as an estimator.

**Example 2 Solutions to Normal Equations** Two generalized inverses of the $\mathbf{X}'\mathbf{X}$ matrix in (6) are

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_2 = \begin{bmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The generalized inverse $\mathbf{G}_1$ is obtained using the inverse of the $3 \times 3$ sub-matrix in the rows and columns 2–4 while $\mathbf{G}_2$ is obtained using the inverse of the $3 \times 3$ sub-matrix in rows and columns 1–3. Then

$$(\mathbf{b}_1^\circ)' = (\mathbf{G}_1\mathbf{X}'\mathbf{y}) = [\,0 \quad 47.51 \quad 0.785 \quad 0.19\,]$$

and

$$(\mathbf{b}_2^\circ)' = (\mathbf{G}_2\mathbf{X}'\mathbf{y}) = [\,0.19 \quad 47.32 \quad 0.595 \quad 0\,]$$

Notice that for both solutions $\mu + \alpha_i$ and $\alpha_i - \alpha_j$, $i, j = 1, 2, 3$ are equal. □

## 2. CONSEQUENCES OF A SOLUTION

The solution $\mathbf{b}^\circ$ to the normal equations is clearly a function of $\mathbf{y}$, the observations, even though it is not an estimator of $\mathbf{b}$. The expected value, the variance, and other results about $\mathbf{b}^\circ$ are therefore not identical to those of $\hat{\mathbf{b}}$ of Chapter 3.

### a. Expected Value of $\mathbf{b}^\circ$

For any generalized inverse $\mathbf{G}$,

$$E(\mathbf{b}^\circ) = \mathbf{G}\mathbf{X}'E(y) = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{b}. \tag{8a}$$

The solution $\mathbf{b}^\circ$ has expected value $\mathbf{H}\mathbf{b}$ where $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$. Thus, $\mathbf{b}^\circ$ is an unbiased estimator of $\mathbf{H}\mathbf{b}$ but not of $\mathbf{b}$.

Consider the solution $\mathbf{b}^{\circ}_{mp} = (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y}$ where $(\mathbf{X}'\mathbf{X})^{+}$ is the Moore–Penrose inverse of $\mathbf{X}'\mathbf{X}$. From Theorem 9 in Chapter 1, $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{X} = \mathbf{U}\Lambda^{-1}\mathbf{U}'\mathbf{U}\Lambda\mathbf{U}' = \mathbf{U}\mathbf{U}'$. Thus, from (8a) we have

$$E(\mathbf{b}^{\circ}_{mp}) = \mathbf{U}\mathbf{U}'\mathbf{b}. \tag{8b}$$

Thus $\mathbf{b}^{\circ}_{mp}$ is an unbiased estimator of $\mathbf{U}\mathbf{U}'\mathbf{b}$ but not of $\mathbf{b}$.

### b. Variance Covariance Matrices of b° (Variance Covariance Matrices)

Let var($\mathbf{b}^{\circ}$) denote the variance covariance matrix of $\mathbf{b}^{\circ}$. Likewise, var($\mathbf{y}$) is the variance covariance matrix for $\mathbf{y}$. From (7),

$$\text{var}(\mathbf{b}^{\circ}) = \text{var}(\mathbf{G}\mathbf{X}'\mathbf{y}) = \mathbf{G}\mathbf{X}'\text{var}(\mathbf{y})\mathbf{X}\mathbf{G} = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma^2. \tag{9}$$

This is not an analogue of its counterpart $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ as would be $\mathbf{G}\sigma^2$.

However, when the choice of $\mathbf{G}$ is such that it is a reflexive generalized inverse, we have that var($\mathbf{b}^{\circ}$) = $\mathbf{G}\sigma^2$ and if $\mathbf{G}$ is the Moore–Penrose inverse $(\mathbf{X}'\mathbf{X})^{+}$ var($\mathbf{b}^{\circ}$) = $(\mathbf{X}'\mathbf{X})^{+}\sigma^2$. A reflexive generalized inverse can always be obtained using the algorithm in Chapter 1. The Moore–Penrose inverse can be obtained using the singular value decomposition as explained in Chapter 1.

**Example 3 Variances of Linear Combinations of Parameters** In Example 2, the individual solutions to the normal equations would have different variances depending on the generalized inverse used to find them. For example, if $\mathbf{G}_1$ was used, then $\alpha^{\circ}_1$ would have variance $(1/3)\sigma^2$, while if $\mathbf{G}_2$ were used, it would have variance $(4/3)\sigma^2$. However, some of the linear combinations (the estimable ones) will have the same variances. For example, consider $\alpha^{\circ}_1 - \alpha^{\circ}_2$. We have using $\mathbf{G}_1$,

$$\text{var}\left(\alpha^{\circ}_1 - \alpha^{\circ}_2\right) = [\,0 \quad 1 \quad -1 \quad 0\,]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}\sigma^2 = \frac{5}{6}\sigma^2.$$

Using $\mathbf{G}_2$, we have

$$\text{var}\left(\alpha^{\circ}_1 - \alpha^{\circ}_2\right) = [\,0 \quad 1 \quad -1 \quad 0\,]\begin{bmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}\sigma^2 = \frac{5}{6}\sigma^2.$$

□

### c. Estimating $E(\mathbf{y})$

Corresponding to the vector of observations $\mathbf{y}$, we have the vector of estimated values $\widehat{E(y)}$, just as in Section 4c of Chapter 3.

$$\widehat{E(\mathbf{y})} = \hat{\mathbf{y}} = \mathbf{X}\mathbf{b}^\circ = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}. \tag{10}$$

This vector is invariant to the choice of the generalized inverse of $\mathbf{X}'\mathbf{X}$ that is used for $\mathbf{G}$ by Theorem 10 in Chapter 1. Hence (10) is *the* vector of estimated expected values corresponding to the vector of observations. This means that no matter what solution of the normal equations is used for $\mathbf{b}^\circ$, the vector $\hat{\mathbf{y}} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}$ will always be the same.

This result and others that will be developed are of great importance. It means that we can get a solution to the normal equations in any way we please, call it $\mathbf{b}^\circ$, and no matter what solution it is, $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}^\circ$ will be the correct value of $\hat{\mathbf{y}}$.

**Example 4 Different Solutions to Normal Equations Predict the Same Values for y** Consider the two solutions $\mathbf{b}_1^\circ$ and $\mathbf{b}_2^\circ$ to the normal equations in Example 2. Observe that

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_1^\circ = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{bmatrix} = \begin{bmatrix} 47.51 \\ 47.51 \\ 47.51 \\ 0.785 \\ 0.785 \\ 0.19 \end{bmatrix}$$

and

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_2^\circ = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 32 \\ 68 \\ 54 \\ 0 \end{bmatrix} = \begin{bmatrix} 47.51 \\ 47.51 \\ 47.51 \\ 0.785 \\ 0.785 \\ 0.19 \end{bmatrix}.$$

□

### d. Residual Error Sum of Squares

As before, the residual error sum of squares is defined as

$$\begin{aligned} \text{SSE} &= (\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ) \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y} \end{aligned} \tag{11}$$

because $\mathbf{I} - \mathbf{XGX}'$ is idempotent and it is symmetric by Theorem 10 of Chapter 1. Furthermore, SSE is invariant to $\mathbf{G}$ because $\mathbf{X}'\mathbf{GX}$ is. Thus, SSE is invariant to whatever solution of the normal equations is used for $\mathbf{b}^\circ$. Thus, equation (11) for SSE is another result invariant to the many solutions there are to the normal equation.

As was the case for the full-rank model, a computing formula for SSE may be derived. Observe that

$$\text{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{XGXy} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ'}\mathbf{X}'\mathbf{y}. \tag{12}$$

This is exactly the same result that was obtained for the full-rank case (recall equation (83) in Section 4d of Chapter 3). We have that $\mathbf{y}'\mathbf{y}$ is the total sum of squares of the observed $y$'s and $\mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ is the sum of the products of the solutions in $\mathbf{b}^{\circ\prime}$, multiplied by the corresponding elements of the right-hand sides of the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ from which $\mathbf{b}^\circ$ is derived.

### e. Estimating the Residual Error Variance

Since $\mathbf{y}$ is distributed with mean $\mathbf{Xb}$ and variance matrix $\sigma^2\mathbf{I}$, equation (45) of Chapter 2 yields

$$E(\text{SSE}) = E[\mathbf{y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{y}] = \text{tr}[(\mathbf{I} - \mathbf{XGX}')\mathbf{I}\sigma^2] + \mathbf{b}'\mathbf{X}'(\mathbf{I} - \mathbf{XGX}')\mathbf{Xb}.$$

Using the properties of $\mathbf{XGX}'$ of Theorem 10, Chapter 1, the above expression reduces to

$$E(\text{SSE}) = \sigma^2 r(\mathbf{I} - \mathbf{XGX}') = [N - r(\mathbf{X})]\sigma^2.$$

Hence an unbiased estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N - r(\mathbf{X})}. \tag{13}$$

We again see a similarity with the full-rank case. However, it is now clear why we should use $r(\mathbf{X})$ in the expectation. Now the matrix $\mathbf{X}$ is not of full-column rank. Therefore, its rank is not equal to the number of columns. In fact, the rank of $\mathbf{X}$ depends on the nature of the data available.

**Example 5 Estimating the Residual Error Variance** We use the data from Examples 1–4. We have that

$$\mathbf{XG}_1\mathbf{X}' = \mathbf{XG}_2\mathbf{X}' = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{I} - \mathbf{XG}_1\mathbf{X}' = \mathbf{I} - \mathbf{XG}_2\mathbf{X}' = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$\mathbf{y}'(\mathbf{I} - \mathbf{XG}_1\mathbf{X}')\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{XG}_2\mathbf{X}')\mathbf{y} = 724.999.$$

Now $r(\mathbf{X}) = 3$ and $N - r(\mathbf{X}) = 3$. Thus,

$$\hat{\sigma}^2 = \frac{724.999}{3} = 241.666.$$

□

## f. Partitioning the Total Sum of Squares

Partitioning the total sum of squares as shown in Section 4f of Chapter 3 for the full-rank model occurs in exactly the same fashion for the non-full-rank model. The only difference is that there is no utility in corrected sums of squares and products of the $x$-variables. Therefore, matrices of the form $\mathcal{X}'\mathcal{X}$ do not arise. However, use is still made of $\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$, the corrected sum of squares of the $y$-observations. The three forms of partitioning the sum of squares are shown in Table 5.3.

The three columns in Table 5.3 correspond to the three partitionings shown in (87), (88), and (90) of Chapter 3. The first column shows

$$\text{SSR} = \text{SST–SSE} = \mathbf{y}'\mathbf{XGX}'\mathbf{y} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}, \tag{14}$$

**TABLE 5.3 Partitioning the Sum of Squares**

| | | |
|---|---|---|
| | $\text{SSM} = N\bar{y}^2 = \mathbf{y}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{y}$ | |
| $\text{SSR} = \mathbf{y}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}$ | $\text{SSR}_\text{m} = \mathbf{y}'(\mathbf{X}\mathbf{G}\mathbf{X}' - N^{-1}\mathbf{1}\mathbf{1}')\mathbf{y}$ | $\text{SSR}_\text{m} = \mathbf{y}'(\mathbf{X}\mathbf{G}\mathbf{X}' - N^{-1}\mathbf{1}\mathbf{1}')\mathbf{y}$ |
| $\text{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}$ | $\text{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}$ | $\text{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}$ |
| $\text{SST} = \mathbf{y}'\mathbf{y}$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | $\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}^2$ |

the sum of squares attributable to fitting the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ similar to the sum of squares due to regression in Chapter 3. In the second column,

$$\text{SSM} = N\bar{y}^2 \tag{15}$$

is the sum of squares due to fitting the general mean and

$$\text{SSR}_\text{m} = \text{SSR–SSM} = \text{SSR–N}\bar{y}^2 \tag{16}$$

is the sum of squares for fitting the model, corrected for the mean. The third column is identical to the second, except that SSM has been deleted from the body of the table and subtracted from SST to give

$$\text{SST}_\text{m} = \text{SST–SSM} = \sum_{i=1}^{N} y_i^2 - N\bar{y}^2 \tag{17}$$

as the total sum of squares corrected for the mean. In all three columns, the residual error sum of squares is the same, SSE of (12).

In Section 3, we will show that Table 5.3 forms the basis of the traditional analysis of variance tables.

### g. Coefficient of Determination

The elements of $\hat{\mathbf{y}}$ given in (10) are the estimated expected values of $y$ corresponding to the observations $\mathbf{y}$. The square of the product moment correlation between the observed $y$'s and the corresponding elements of $\hat{\mathbf{y}}$ is commonly referred to as the coefficient of determination. Since the usual linear model has a mean in it, we define

$$\begin{aligned} R^2 &= \text{coefficient of determination} \\ &= \frac{\left[\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}. \end{aligned} \tag{18}$$

To simplify (18), we utilize $\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}'$ (Theorem 10, Chapter 1). The first row of $\mathbf{X}'$ is $\mathbf{1}'$. Thus,

$$1'\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{1}'. \tag{19}$$

As a result,

$$\bar{\hat{y}} = N^{-1}\mathbf{1}\hat{\mathbf{y}} = N^{-1}\mathbf{1}'\mathbf{X}\mathbf{b}^{\circ} = N^{-1}\mathbf{1}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} = N^{-1}\mathbf{1}'\mathbf{y} = \bar{y}.$$

Hence as in equations (91) of Chapter 3,

$$R^2 = \frac{(\mathrm{SSR_m})^2}{\mathrm{SST_m}(\mathrm{SSR_m})} = \frac{\mathrm{SSR_m}}{\mathrm{SST_m}}. \tag{20}$$

The expression in (20) above represents the proportion of the variation that is accounted for by the regression model.

**Example 6 Sums of Squares in Examples 1–5** We have that either

$$\mathrm{SSR} = \mathbf{b}_1^{\circ'}\mathbf{X}'\mathbf{y} = (47.51)(142.53) + (0.785)(1.57) + (0.19)(0.19) = 6772.87 \tag{21}$$

or

$$\mathrm{SSR} = \mathbf{b}_2'^{\circ}\mathbf{X}'\mathbf{y} = (0.19)(144.29) + (47.32)(142.53) + (1.57)(0.595) = 6772.87. \tag{22}$$

Notice that the use of both solutions of the normal equations derived from different generalized inverses gives the same result for the sums of squares. Recall that the vector of observations is

$$\mathbf{y}' = [\,27 \quad 50.91 \quad 64.62 \quad 0.67 \quad 0.9 \quad 0.19\,].$$

Thus,

$$\mathrm{SST} = \sum_{i=1}^{6} y_i^2 = \mathbf{y}'\mathbf{y} = 7497.87 \tag{23}$$

and

$$\mathrm{SSM} = N\bar{y}^2 = 3469.93 \tag{24}$$

Hence the partitioning of sums of squares shown in Table 5.3 is, for the example as given in Table 5.4.

The value of $R^2$, calculated from (20), is $R^2 = 3302.94/4027.94 = 0.82$.

**TABLE 5.4 Partitioning Sums of Squares (Data of Table 5.1)**

| | | |
|---|---|---|
| | SSM = 3469.93 | |
| SSR = 6772.87 | $SSR_m$ = 3302.94 | $SSR_m$ = 3302.94 |
| SSE = 724.999 | SSE = 724.999 | SSE = 724.999 |
| SST = 7497.87 | SST = 7497.87 | $SST_m$ = 4027.94 |

□

## 3. DISTRIBUTIONAL PROPERTIES

We now assume normality for the error terms. Thus, we have that

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N).$$

Using the normality assumption, we shall derive the distributional properties of $\mathbf{y}$ in a manner similar to the full-rank case (see Section 5 of Chapter 3).

### a. The Observation Vector y is Normal

Since $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ and $E(\mathbf{y}) = \mathbf{Xb}$, we have that $\mathbf{y} \sim N(\mathbf{Xb}, \sigma^2\mathbf{I})$.

### b. The Solution to the Normal Equations b° is Normally Distributed

Since $\mathbf{b}^\circ$ is a linear function of $\mathbf{y}$, it is normally distributed with mean and variance derived in (8) and (9). Thus,

$$\mathbf{b}^\circ = \mathbf{GX}'\mathbf{y} \sim N(\mathbf{Hb}, \mathbf{GX}'\mathbf{XG}\sigma^2).$$

Notice that the covariance matrix of $\mathbf{b}^\circ$ is singular.

### c. The Solution to the Normal Equations b° and the Estimator of the Residual Error Variance $\hat{\sigma}^2$ are Independent

We apply Theorem 6 of Chapter 2 to

$$\mathbf{b} = \mathbf{GX}'\mathbf{y} \quad \text{and} \quad \text{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{y}.$$

We see that, $\mathbf{GX}'\mathbf{I}\sigma^2(\mathbf{I} - \mathbf{XGX}') = \mathbf{G}(\mathbf{X}' - \mathbf{X}'\mathbf{XGX}')\sigma^2 = \mathbf{0}$ because $\mathbf{X}' = \mathbf{X}'\mathbf{XGX}'$ (Theorem 10, Chapter 1). Therefore, $\mathbf{b}^\circ$ and $\hat{\sigma}^2$ are independent.

### d. The Error Sum of Squares Divided by the Population Variance SSE/$\sigma^2$ is Chi-square $\chi^2$

We have that

$$\frac{\text{SSE}}{\sigma^2} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{y}}{\sigma^2}.$$

Applying Theorem 5 of Chapter 2, we have

$$\frac{\mathbf{I}\sigma^2(\mathbf{I}-\mathbf{XGX}')}{\sigma^2} = \mathbf{I}-\mathbf{XGX}',$$

an idempotent matrix. Therefore, by Theorem 5 of Chapter 2,

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^{2'}\left[r(\mathbf{I}-\mathbf{XGX}'), \frac{\mathbf{b}'\mathbf{X}'(\mathbf{I}-\mathbf{XGX}')\mathbf{Xb}}{2\sigma^2}\right].$$

Using Theorem 10 of Chapter 1 and $r(\mathbf{X}) = \text{r}$ this reduces to

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{N-r}. \tag{25}$$

### e. Non-central $\chi^{2'}$s

With $\text{SSE}/\sigma^2$ being $\chi^2_{N-r}$, we now show that the other terms in Table 5.3 have non-central $\chi^2$-distributions independent of SSE. This leads to $F$-statistics that have non-central $F$-distributions that, in turn, are central $F$-distributions under certain null hypothesis. Tests of hypothesis are thus established.

From (14), $\text{SSR} = \mathbf{y}'\mathbf{XGX}'\mathbf{y}$. The matrix $\mathbf{XGX}'$ is idempotent and its products with $\mathbf{I}-\mathbf{XGX}'$ are null. Therefore, by Theorems 5 and 7 of Chapter 2, $\text{SSR}/\sigma^2$ is distributed independently of SSE with

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^{2'}\left[r(\mathbf{XGX}'), \frac{\mathbf{b}'\mathbf{X}'\mathbf{XGX}'\mathbf{Xb}}{2\sigma^2}\right] \sim \chi^{2'}\left(r, \frac{\mathbf{b}'\mathbf{X}'\mathbf{Xb}}{2\sigma^2}\right). \tag{26}$$

Similarly,

$$\frac{\text{SSM}}{\sigma^2} = \frac{\mathbf{y}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{y}}{\sigma^2}$$

where $N^{-1}\mathbf{1}\mathbf{1}'$ is idempotent. From (19),

$$N^{-1}\mathbf{1}\mathbf{1}'\mathbf{XGX}' = N^{-1}\mathbf{1}\mathbf{1}'. \tag{27}$$

Thus, the products of $N^{-1}\mathbf{1}\mathbf{1}'$ and $(\mathbf{I}-\mathbf{XGX}')$ are null. Hence, SSM is distributed independently of SSE and

$$\frac{\text{SSM}}{\sigma^2} \sim \chi^{2'}\left[r(N^{-1}\mathbf{1}\mathbf{1}'), \frac{\mathbf{b}'\mathbf{X}'N^{-1}\mathbf{1}\mathbf{1}'\mathbf{Xb}}{2\sigma^2}\right] \sim \chi^{2'}\left(r, \frac{\mathbf{b}'\mathbf{X}'\mathbf{Xb}}{2\sigma^2}\right), \tag{28}$$

just as in the full-rank case.

A similar argument applies to

$$\mathrm{SSR_m} = \mathbf{y}'(\mathbf{XGX}' - N^{-1}\mathbf{11}')\mathbf{y}.$$

The matrix $\mathbf{XGX}' - N^{-1}\mathbf{11}'$ is idempotent (using (27)) and has null products with both $N^{-1}\mathbf{11}'$ and $(\mathbf{I} - \mathbf{XGX}')$. Hence, by Theorems 5 and 7 of Chapter 2, $\mathrm{SSR_m}$ is independent of SSM and SSE. Furthermore,

$$\begin{aligned}\frac{\mathrm{SSR_m}}{\sigma^2} &\sim \chi^{2'}\left[r(\mathbf{XGX}' - N^{-1}\mathbf{11}'), \frac{\mathbf{b}'\mathbf{X}'(\mathbf{XGX}' - N^{-1}\mathbf{11}')\mathbf{Xb}}{2\sigma^2}\right] \\ &\sim \chi^{2'}\left[r-1, \frac{\mathbf{b}'\mathbf{X}'(\mathbf{I} - N^{-1}\mathbf{11}')\mathbf{Xb}}{2\sigma^2}\right].\end{aligned} \tag{29}$$

Now, for any $\mathbf{X}$ whose first column is $\mathbf{1}$ we can write $\mathbf{X} = [\,\mathbf{1}\ \ \mathbf{X}_1\,]$. Then

$$\mathbf{X}'(\mathbf{I} - N^{-1}\mathbf{11}')\mathbf{X} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_1'(\mathbf{I} - N^{-1}\mathbf{11}')\mathbf{X}_1 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathcal{X}'\mathcal{X} \end{bmatrix}. \tag{30}$$

The matrix $\mathcal{X}'\mathcal{X}$ is the same as that defined in Chapter 3. It represents the sums of squares and products of the deviations of elements of the columns (other than the first column) of $\mathbf{X}$ from their means. Symbolically, it is simpler than its equivalent form $\mathbf{X}_1'(\mathbf{I} - N^{-1}\mathbf{11}')\mathbf{X}_1$ but offers little advantage computationally, in distinction to the full-rank model where it is advantageous. Nevertheless, writing

$$\mathbf{b} = \begin{bmatrix} b_0 \\ \boldsymbol{\beta} \end{bmatrix} \tag{31}$$

just as in the full-rank case with $b_0$ representing a general mean, we have from (29) and (30),

$$\frac{\mathrm{SSR_m}}{\sigma^2} \sim \chi^{2'}\left[r-1, \frac{\boldsymbol{\beta}'\mathbf{X}_1'(\mathbf{I} - N^{-1}\mathbf{11}')\mathbf{X}_1\boldsymbol{\beta}}{2\sigma^2}\right] \sim \chi^{2'}\left[\frac{\mathbf{b}'\mathcal{X}'\mathcal{X}\boldsymbol{\beta}}{2\sigma^2}\right]. \tag{32}$$

### f. Non-central *F*-distributions

We obtain the following $F$-statistics from the results in (26), (28), and (35) using the definition of the non-central $F$-distribution. Recall that random variables that follow a non-central $F$-distribution are the ratio of random variables that follow a non-central chi-square distribution and a central chi-square distribution. We find that

$$F(R) = \frac{\mathrm{SSR}/r}{\mathrm{SSE}/(N-r)} \sim F'\left(r, N-r, \frac{\mathbf{b}'\mathbf{X}'\mathbf{Xb}}{2\sigma^2}\right), \tag{33}$$

$$F(M) = \frac{\text{SSM}/1}{\text{SSE}/(N-r)} \sim F'\left[1, N-r, \frac{(\mathbf{1}'\mathbf{X}\mathbf{b})^2}{2N\sigma^2}\right] \tag{34}$$

and

$$F(R_{\text{m}}) = \frac{\text{SSR}_{\text{m}}/(r-1)}{\text{SSE}/(N-r)} \sim F'(r-1, N-r, \mathscr{b}\mathscr{X}'\mathscr{X}\mathscr{b}). \tag{35}$$

Under certain null hypotheses, these non-central $F$'s become central $F$'s and so provide us with tests of hypothesis. We shall discuss these in Section 3h and in Section 5.

## g. Analyses of Variance

Calculation of the above $F$-statistics can be summarized in analysis of variance tables just as was done in Tables 5.2, 5.3, and 5.4 of Section 5h of Chapter 3. The sums of squares are those of Table 5.3.

Table 5.5 summarizes not only the sums of squares but also the degrees of freedom associated with the $\chi^2$-distributions. It also shows, in the mean squares, the calculation of the numerator and denominator of $F(R)$ of equation (33) as well as $F(R)$ itself. Therefore, the table is a convenient summary of the calculations.

Table 5.6 shows the same thing for $F(M)$ and $F(R)$ of (34) and (35). Table 5.6b shows the abbreviated form of the complete analysis of variance table shown in Table 5.6a. The derivation of this abbreviated form consists of removing SSM from the body of the table and subtracting it from SST to give $\text{SST}_{\text{m}}$ as in Table 5.3. Thus, Table 5.6b does not contain $F(M)$. However, it is identical to Table 5.6a as far as $F(R_{\text{m}}) = \text{MSR}_{\text{m}}/\text{MSE}$ is concerned. Thus, the two sections of Table 5.6 are similar to Tables 3.3 and 3.4 of Section 5h of Chapter 3.

Although Table 5.6b is the form in which this analysis of variance is most usually seen, it is not the most informative. Table 5.6a has more information because it shows how SSR of Table 5.4 gets partitioned into SSM and $\text{SSR}_{\text{m}}$, and thus summarizes $F(M)$ and $F(R_{\text{m}})$.

**TABLE 5.5 Analysis of Variance for Fitting the Model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Model | $r = r(\mathbf{X})$ | $\text{SSR} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ | $\text{MSR} = \dfrac{\text{SSR}}{r}$ | $F(R) = \dfrac{\text{MSR}}{\text{MSE}}$ |
| Residual Error | $N - r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| Total | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | | |

**TABLE 5.6 Analysis of Variance for Fitting the Model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$**

(a) Complete form

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistics |
|---|---|---|---|---|
| Mean | 1 | $\text{SSM} = N\bar{y}^2$ | $\text{MSM} = \dfrac{\text{SSM}}{1}$ | $F(M) = \dfrac{\text{MSM}}{\text{MSE}}$ |
| Model (a.f.m.) | $r-1$ | $\text{SSR}_\text{m} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} - N\bar{y}^2$ | $\text{MSR}_\text{m} = \dfrac{\text{SSR}_\text{m}}{r-1}$ | $F(\text{R}_\text{m}) = \dfrac{\text{MSR}_\text{m}}{\text{MSE}}$ |
| Residual error | $N-r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| Total | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | | |

a.f.m. = after fitting the mean. Also $r = r(\mathbf{X})$.

(b) Abbreviated form

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistics |
|---|---|---|---|---|
| Model (a.f.m.) | $r-1$ | $\text{SSR}_\text{m} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} - N\bar{y}^2$ | $\text{MSR}_\text{m} = \dfrac{\text{SSR}_\text{m}}{r-1}$ | $F(R_\text{m}) = \dfrac{\text{MSR}_\text{m}}{\text{SSE}}$ |
| Residual error | $N-r$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ | $\text{MSE} = \dfrac{\text{SSE}}{N-r}$ | |
| Total | $N-1$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ | | |

### h. Tests of Hypotheses

The results in equations (33)–(35) provide statistics suitable for testing certain hypotheses. We will now discuss this. Later in Section 5, we will consider the general linear hypothesis. The general linear hypothesis will contain the results discussed here as special cases.

The $F(R)$-statistic of (33), whose calculation is summarized in Table 5.5, has a non-central $F$-distribution with non-centrality parameter $\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/2\sigma^2$. This non-centrality parameter is zero under the null hypothesis $H$: $\mathbf{Xb} = \mathbf{0}$. The statistic $F(R)$ then has a central $F_{r,N-r}$-distribution. The calculated value of the $F$-statistic can be compared with tabulated values to test the null hypothesis $H$. When $F(R)$ is significant, we might say just as we did in Section 5h of Chapter 3 that there is concordance if the data with the model $E(y) = \mathbf{Xb}$. In other words, the model accounts for a significant portion of the variation in the $y$ variable. This does not mean that the model used is necessarily the most suitable model. The following are possible contingencies.

(i) There may be a subset of the elements that is as significant as the whole set.

(ii) There may be other elements (factors) which, when used alone, or in combination with some or all of those already used, are significantly better than those used.

(iii) There may be nonlinear models that are at least as suitable as the model used.

None of the above contingencies is inconsistent with $F(R)$ being significant and the ensuing conclusion that the data are in concordance with the model $E(\mathbf{y}) = \mathbf{Xb}$.

Notice that, in contrast to the full-rank case in Section 5h of Chapter 3 that the test based on $F(R)$ cannot be described formally as testing $H$: $\mathbf{b} = \mathbf{0}$ because as we shall show in Sections 4 and 5 that $\mathbf{b}$ is not what we call an "estimable function". This means $H$: $\mathbf{b} = \mathbf{0}$ cannot be tested, but $H$: $\mathbf{Xb} = \mathbf{0}$ can be tested. We will soon show that $F(R)$ is the appropriate statistic.

The non-centrality parameter of $F(M)$ in Table 5.6a is, by (34), $(\mathbf{1}'\mathbf{Xb})^2/2N\sigma^2$. Just as in the full-rank case (Section 5h of Chapter 3) this parameter equals $N[E(\bar{y})]^2/2\sigma^2$. Under the null hypothesis, $H$: $E(\bar{y}) = 0$ it is zero. Then, the statistic $F(M)$ is distributed as $F_{1,N-r}$. Hence, $F(M)$ provides a test of the hypothesis $H$: $E(\bar{y}) = 0$. The test is based on comparing $F(M)$ with the tabulated values of the $F_{1,N-r}$-distribution. An equivalent test is to compare $\sqrt{F(M)}$ against tabulations of the $t$-distribution having $N - r$ degrees of freedom. This hypothesis $H$: $E(\bar{y}) = 0$ is one interpretation of what is meant by "testing the mean". Another interpretation, just as in the full-rank case is that $F(M)$ can be used to test whether the model $E(y_{ij}) = b_0$ accounts for variation in the $y$-variable.

Just as $F(R)$ provides a test of the model $E(\mathbf{y}) = \mathbf{Xb}$, so does $F(R_{\mathrm{m}})$ provide a test of the model over and above the mean. For the same reason that $F(R)$ cannot be described as testing $H$: $\mathbf{b} = \mathbf{0}$, also $F(R_{\mathrm{m}})$ cannot be described as testing $H$: $\boldsymbol{\mathscr{b}} = \mathbf{0}$. Again $\boldsymbol{\mathscr{b}}$ is not, in general, what is called an "estimable function" and so $H$: $\boldsymbol{\mathscr{b}} = \mathbf{0}$ cannot be tested (see Sections 4 and 5). In general, therefore, $F(R_{\mathrm{m}})$ must be looked upon as providing a test of the model $E(\mathbf{y}) = \mathbf{Xb}$ over and above the mean. When $F(R_{\mathrm{m}})$ is significant, we conclude that the model significantly accounts for the variation in the $y$-variable. This is not to be taken as evidence that all of the elements of $\mathbf{b}$ are non-zero, but only that at least one of them, or one linear combination of them, may be. If $F(M)$ has first been found significant, then $F(R_{\mathrm{m}})$ being significant indicates that a model with terms in it additional to a mean explains significantly more of the variation in the $y$-variable than does the model $E(y) = b_0$.

Similar to regression, the tests using $F(M)$ and $F(R_{\mathrm{m}})$ are based on numerators that are statistically independent although their denominators, the residual mean square, are identical. The $F$-statistics are therefore not independent.

The case of both $F(M)$ and $F(R_{\mathrm{m}})$ being significant has just been discussed and illustrated in Table 5.6a. Another possibility is that $F(M)$ is not significant but $F(R_{\mathrm{m}})$ is. This would be evidence that the mean is zero, but that fitting the rest of the model explains variation in the $y$ variable. As in regression, a likely situation when this might occur is when the $y$ variable can have both positive and negative values.

**Example 7 Interpretation of the Results in Table 5.7** The analysis of variance of Tables 5.5 and 5.6 are shown in Table 5.7. They use the sums of squares in Table 5.4. In this case, all three $F$-statistics $F(R)$, $F(M)$, and $F(R_{\mathrm{m}})$ are significant. This indicates respectively that:

(i) The model almost accounts for a significant portion of the variation in $y$.
(ii) The mean is unlikely to be zero.
(iii) The model needs something more than the mean to explain variation in $y$.

**TABLE 5.7 Tables 5.5 and 5.6 for Example 6**

| Source of Variation | d.f | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| | | Table 5.5 | | |
| Model | 3 | SSR = 6772.87 | 2257.62 | $F(R) = 9.3419$ |
| Residual error | 3 | SSE = 724.999 | 241.666 | |
| Total | 6 | SST = 7497.87 | | |
| | | Table 5.6a | | |
| Mean | 1 | SSM = 3469.93 | 3469.93 | $F(M) = 14.3584$ |
| Model | 2 | $SSR_m = 3{,}302.94$ | 1651.47 | $F(R_m) = 6.8337$ |
| Residual error | 3 | SSE = 724.999 | 241.666 | |
| Total | 6 | SST = 7497.87 | | |
| | | Table 5.6b | | |
| Model (a.f.m.) | 2 | $SSR_m = 3{,}302.94$ | 1651.47 | $F(R_m) = 6.8337$ |
| Residual error | 3 | SSE = 724.999 | 241.666 | |
| Total | 5 | $SST_m = 4027.94$ | | |

□

## 4. ESTIMABLE FUNCTIONS

The underlying idea of an estimable function was introduced at the end of Chapter 4. Basically, it is a linear function of the parameters for which an estimator can be found from $\mathbf{b}^\circ$ that is invariant to whatever solution to the normal equations that is used for $\mathbf{b}^\circ$. There were a number of exercises and examples that illustrated this property. We will not discuss such functions in detail. We confine ourselves to linear functions of the form $\mathbf{q}'\mathbf{b}$ where $\mathbf{q}'$ is a row vector.

### a. Definition

A linear function of the parameters is defined as *estimable* if it is identically equal to some linear function of the expected value of the vector of observations. This means that $\mathbf{q}'\mathbf{b}$ is estimable if $\mathbf{q}'\mathbf{b} = \mathbf{t}'E(\mathbf{y})$ for some vector $\mathbf{t}'$. In other words, if a vector $\mathbf{t}'$ exists such that $\mathbf{t}'E(\mathbf{y}) = \mathbf{q}'\mathbf{b}$, then $\mathbf{q}'\mathbf{b}$ is said to be estimable. Note that in no way is there any sense of uniqueness about $\mathbf{t}'$. It simply has to exist.

**Example 8 An Estimable Function** Consider the model used for Examples 1–7. Consider the function $\alpha_1 - \alpha_2$. We have that $\mathbf{q}' = [\,0 \;\; 1 \;\; -1 \;\; 0\,]$. Then,

$$\begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} = [\,t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_5\,] \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

because $\mathbf{q}'\mathbf{b} = E(\mathbf{t}'\mathbf{y}) = \mathbf{t}'\mathbf{X}\mathbf{b}$ for all $\mathbf{b}$ so the condition for a linear function to be estimable reduces to the existence of a $\mathbf{t}$, where $\mathbf{q}' = \mathbf{t}'\mathbf{X}$. The system of equations

$$\begin{aligned} 0 &= t_1 + t_2 + t_3 + t_4 + t_5 + t_6 \\ 1 &= t_1 + t_2 + t_3 \\ -1 &= t_4 + t_5 \\ 0 &= t_6 \end{aligned}$$

has infinitely many solutions. Two of them are $t_1 = t_2 = t_3 = \frac{1}{3}, t_4 = t_5 = -\frac{1}{2}, t_6 = 0$ and $t_1 = t_2 = \frac{1}{4}, t_3 = \frac{1}{2}, t_4 = -\frac{2}{3}, t_5 = -\frac{1}{3}, t_6 = 0$. On the other hand, the system of equations

$$\begin{aligned} 0 &= t_1 + t_2 + t_3 + t_4 + t_5 + t_6 \\ 1 &= t_1 + t_2 + t_3 \\ 0 &= t_4 + t_5 \\ 0 &= t_6 \end{aligned}$$

is inconsistent and has no solution. Therefore, the individual parameter $\alpha_1$ is not estimable.

The value of $\mathbf{t}'$ is not as important as its existence. In this sense, all that needs to be done to establish estimability of $\mathbf{q}'\mathbf{b}$ is to be satisfied that there is at least one linear function of the expected values of the $y$'s, $\mathbf{t}'E(\mathbf{y})$, whose value is $\mathbf{q}'\mathbf{b}$. Since $\mathbf{t}'E(\mathbf{y}) = E(\mathbf{t}'\mathbf{y})$, this is equivalent to establishing some linear function of the $y$'s, $\mathbf{t}'\mathbf{y}$, whose expected value is $\mathbf{q}'\mathbf{b}$. There are usually many such functions of the $y$'s. Establishing existence of any one of them is sufficient for establishing estimability.

### b. Properties of Estimable Functions

**(*i*) *The Expected Value of Any Observation is Estimable*** The definition of an estimable function is that $\mathbf{q}'\mathbf{b}$ is estimable if $\mathbf{q}'\mathbf{b} = \mathbf{t}'E(\mathbf{y})$ for some vector $\mathbf{t}'$. Consider a $\mathbf{t}'$ which has one element unity and the others zero. Then, $\mathbf{t}'E(\mathbf{y})$ will be estimable. It is an element of $E(\mathbf{y})$, the expected value of an observation. Hence, the expected value of any observation is estimable. For example, for the linear model of Examples 1–8, $E(y_{1j}) = \mu + \alpha_1$ and so $\mu + \alpha_1$ is estimable.

**(*ii*) *Linear Combinations of Estimable Functions are Estimable*** Every estimable function is a linear combination of the elements of $E(\mathbf{y})$. This is also true about a linear combination of estimable functions. Thus, a linear combination of estimable functions is also estimable. More formally, if $\mathbf{q}_1'\mathbf{b}$ and $\mathbf{q}_2'\mathbf{b}$ are estimable, there exists a $\mathbf{t}_1'$ and a $\mathbf{t}_2'$ such that $\mathbf{q}_1'\mathbf{b} = \mathbf{t}_1'E(\mathbf{y})$ and $\mathbf{q}_2'\mathbf{b} = \mathbf{t}_2'E(\mathbf{y})$. Hence, a linear combination $c_1\mathbf{q}_1'\mathbf{b} + c_2\mathbf{q}_2'\mathbf{b} = (c_1\mathbf{t}_1' + c_2\mathbf{t}_2')E(\mathbf{y})$ and so it is estimable.

(*iii*) ***The Forms of an Estimable Function*** If $\mathbf{q}'\mathbf{b}$ is estimable using its definition, we have that for some vector $\mathbf{t}'$

$$\mathbf{q}'\mathbf{b} = \mathbf{t}'E(\mathbf{y}) = \mathbf{t}'E(\mathbf{X}\mathbf{b}) = \mathbf{t}'\mathbf{X}\mathbf{b}. \tag{36}$$

Since estimability is a concept that does not depend on the value of $\mathbf{b}$, the result in (36) must be true for all $\mathbf{b}$. Therefore,

$$\mathbf{q}' = \mathbf{t}'\mathbf{X} \tag{37}$$

for some vector $\mathbf{t}'$. This is equivalent to saying that $\mathbf{q}$ is in the column space of $\mathbf{X}$, the vector space generated by the linear combinations of the columns of $\mathbf{X}$. For any estimable function $\mathbf{q}'\mathbf{b}$, the specific value of $\mathbf{t}'$ is unimportant. What is important is the *existence* of some $\mathbf{t}'$. We shall use (37) repeatedly in the sequel. We have that $\mathbf{q}'\mathbf{b}$ is estimable whenever $\mathbf{q}' = \mathbf{t}'\mathbf{X}$. Conversely, estimability of $\mathbf{q}'\mathbf{b}$ implies $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ for some $\mathbf{t}'$.

Another equivalent condition for estimability $\mathbf{q}'\mathbf{b}$ is that there exists a vector $\mathbf{d}'$ such that

$$\mathbf{q}' = \mathbf{d}'\mathbf{U}' \tag{38}$$

where $\mathbf{U}$ is from the singular value decomposition of $\mathbf{X}$, $\mathbf{X} = \mathbf{S}'\Lambda^{1/2}\mathbf{U}'$. From (37), we have that $\mathbf{q}' = \mathbf{t}'\mathbf{X} = \mathbf{t}'\mathbf{S}'\Lambda^{1/2}\mathbf{U}'$. Thus, the existence of $\mathbf{t}'$ implies the existence of $\mathbf{d}'$ because $\mathbf{d}' = \mathbf{t}'\mathbf{S}'\Lambda^{1/2}$. On the other hand, if (38) holds true since $\mathbf{X} = \mathbf{S}'\Lambda^{1/2}\mathbf{U}'$, $\mathbf{U}' = \Lambda^{-1/2}\mathbf{S}\mathbf{X}$, so $\mathbf{t}' = \mathbf{d}'\Lambda^{-1/2}\mathbf{S}$. Thus, existence of a $\mathbf{d}'$ implies existence of a $\mathbf{t}'$.

(*iv*) ***Invariance to the Solution*** $\mathbf{b}^\circ$ When $\mathbf{q}'\mathbf{b}$ is estimable, $\mathbf{q}'\mathbf{b}^\circ$ is invariant to whatever solution of the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ is used for $\mathbf{b}^\circ$. This is true because from (37), $\mathbf{q}'\mathbf{b}^\circ = \mathbf{t}'\mathbf{X}\mathbf{b}^\circ = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}$ and $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to $\mathbf{G}$ (Theorem 10, Chapter 1). Therefore, $\mathbf{q}'\mathbf{b}^\circ$ is invariant to $\mathbf{G}$ and hence to $\mathbf{b}^\circ$, when $\mathbf{q}'\mathbf{b}$ is estimable. This is why estimability is very important. If $\mathbf{q}'\mathbf{b}$ is estimable, then $\mathbf{q}'\mathbf{b}^\circ$ has the same value for all solutions $\mathbf{b}^\circ$ to the normal equations.

Alternatively, using the singular value decomposition, (38), and Theorem 9 from Chapter 1, we have that

$$\mathbf{q}'\mathbf{b}^\circ = \mathbf{d}'\mathbf{U}'\mathbf{G}\mathbf{X}'\mathbf{y} = \mathbf{d}'\mathbf{U}'\mathbf{G}\mathbf{U}\Lambda^{1/2}\mathbf{S}\mathbf{y} = \mathbf{d}'\Lambda^{-1/2}\Lambda^{1/2}\mathbf{S}\mathbf{y} = \mathbf{d}'\mathbf{S}\mathbf{y}. \tag{39}$$

The last expression in (39) is invariant to $\mathbf{G}$.

(*v*) ***The Best Linear Unbiased Estimator (b.l.u.e.) Gauss–Markov Theorem*** In Chapter 3, we established for the full-rank model, that the least-square estimator was the best linear unbiased estimator of the parameters $\mathbf{b}$ in the regression model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$. We now establish that for estimable linear combinations of the parameters, the estimable linear combinations of solutions to the normal are best linear unbiased estimators for the less than full-rank case.

**Theorem 1** (Gauss–Markov Theorem). The best linear unbiased estimator of the estimable function $\mathbf{q}'\mathbf{b}$ is $\mathbf{q}'\mathbf{b}^\circ$; that is,

$$\widehat{\mathbf{q}'\mathbf{b}} = \mathbf{q}'\mathbf{b}^\circ, \tag{40}$$

where by the "hat" notation we mean "b.l.u.e. of".

*Proof.* To establish (41), we demonstrate properties of linearity, unbiasedness, and "bestness" (having minimum variance). First, $\mathbf{q}'\mathbf{b}^\circ$ is a linear function of the observations, because $\mathbf{q}'\mathbf{b}^\circ = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{y}$. Second, $\mathbf{q}'\mathbf{b}^\circ$ is an unbiased estimator of $\mathbf{q}'\mathbf{b}$ because

$$E(\mathbf{q}'\mathbf{b}^\circ) = \mathbf{q}'E(\mathbf{b}^\circ) = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{t}'\mathbf{X}\mathbf{b} = \mathbf{q}'\mathbf{b}. \tag{41}$$

In establishing (41), we invoke (39) and from Theorem 10 of Chapter 1,

$$\mathbf{X} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} \text{ which also implies } \mathbf{X}' = \mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}. \tag{42}$$

Alternatively, using Theorem 9 from Chapter 1,

$$E(\mathbf{q}'\mathbf{b}^\circ) = \mathbf{q}'E(\mathbf{b}^\circ) = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{d}'\mathbf{U}'\mathbf{G}\mathbf{U}\Lambda\mathbf{U}'\mathbf{b} = \mathbf{d}'\Lambda^{-1}\Lambda\mathbf{U}'\mathbf{b} = \mathbf{d}'\mathbf{U}'\mathbf{b} = \mathbf{q}'\mathbf{b}.$$

To demonstrate that $\mathbf{q}'\mathbf{b}^\circ$ is a best estimator, we need its variance. We then show that the variance of any other linear unbiased estimator of $\mathbf{q}'\mathbf{b}$ is larger. We have that

$$\begin{aligned} v(\mathbf{q}'\mathbf{b}^\circ) &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{q}\sigma^2 && \text{form (9)} \\ &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{t}\sigma^2 && \text{form (37)} \\ &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{t}\sigma^2 && \text{form (42)} \\ &= \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2 && \text{form (37).} \end{aligned} \tag{43}$$

Using the result derived in (43), following C.R. Rao (1962), we now show that $\mathbf{q}'\mathbf{b}^\circ$ has the minimum variance among all the linear unbiased estimators $\mathbf{q}'\mathbf{b}$ and hence is the best. Suppose that $\mathbf{k}'\mathbf{y}$ is some other linear unbiased estimator of $\mathbf{q}'\mathbf{b}$ different from $\mathbf{q}'\mathbf{b}^\circ$. Then, because $\mathbf{k}'\mathbf{y}$ is unbiased, $E(\mathbf{k}'\mathbf{y}) = \mathbf{q}'\mathbf{b}$ so $\mathbf{k}'\mathbf{X} = \mathbf{q}'$. Therefore,

$$\operatorname{cov}(\mathbf{q}'\mathbf{b}^\circ, \mathbf{k}'\mathbf{y}) = \operatorname{cov}(\mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{y}, \mathbf{k}'\mathbf{y}) = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{k}\sigma^2 = \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2.$$

Consequently,

$$\begin{aligned} v(\mathbf{q}'\mathbf{b}^\circ - \mathbf{k}'\mathbf{y}) &= v(\mathbf{q}'\mathbf{b}^\circ) + v(\mathbf{k}'\mathbf{y}) - 2\operatorname{cov}(\mathbf{q}'\mathbf{b}^\circ, \mathbf{k}'\mathbf{y}) \\ &= v(\mathbf{k}'\mathbf{y}) - \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2 \\ &= v(\mathbf{k}'\mathbf{y}) - v(\mathbf{q}'\mathbf{b}^\circ) > 0. \end{aligned} \tag{44}$$

Since $v(\mathbf{q}'\mathbf{b}^\circ - \mathbf{k}'\mathbf{y})$ is positive, from (44) $v(\mathbf{k}'\mathbf{y})$ exceeds $v(\mathbf{q}'\mathbf{b}^\circ)$. Thus, $\mathbf{q}'\mathbf{b}^\circ$ has a smaller variance than any other linear unbiased estimator of $\mathbf{q}'\mathbf{b}$ and hence is the best. ■

The importance of this result must not be overlooked. If $\mathbf{q}'\mathbf{b}$ is an estimable function, its b.l.u.e. is $\mathbf{q}'\mathbf{b}^\circ$ with variance $\mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2$. This is so for any solution $\mathbf{b}^\circ$ to the normal equations using any generalized inverse $\mathbf{G}$. Both the estimator and its variance are invariant to the choice of $\mathbf{G}$ and $\mathbf{b}^\circ$. However, this is true *only* for estimable functions and not for non-estimable functions.

The covariance between the b.l.u.e.'s of two estimable functions is derived in a manner similar to (43). It is

$$\operatorname{cov}(\mathbf{q}_1'\mathbf{b}^\circ, \mathbf{q}_2'\mathbf{b}^\circ) = \mathbf{q}_1'\mathbf{G}\mathbf{q}_2\sigma^2. \tag{45}$$

Hence if $\mathbf{Q}'\mathbf{b}^\circ$ represents the b.l.u.e.'s of several estimable functions, the variance–covariance matrix of these b.l.u.e.'s is

$$\operatorname{var}(\mathbf{Q}'\mathbf{b}^\circ) = \mathbf{Q}'\mathbf{G}\mathbf{Q}\sigma^2. \tag{46}$$

### c. Confidence Intervals

The establishment of confidence intervals is only valid for estimable functions because they are the only functions that have estimators (b.l.u.e.'s) invariant to the solution to the normal equations. Similar to equation (108) of Chapter 3, we have, on the basis of normality, that the symmetric $100(1-\alpha)\%$ confidence interval on the estimable function $\mathbf{q}'\mathbf{b}$ is

$$\mathbf{q}'\mathbf{b}^\circ \pm \hat{\sigma} t_{N-r,\frac{1}{2}\alpha} \sqrt{\mathbf{q}'\mathbf{G}\mathbf{q}}. \tag{47}$$

The probability statement $\Pr\{t \geq t_{N-r,\frac{1}{2}\alpha}\} = \frac{1}{2}\alpha$ defines the value $t_{N-r,\frac{1}{2}\alpha}$ for $t$ having the $t$-distribution with $N-r$ degrees of freedom. As before, when $N-r$ is large ($N-r \geq 100$, say), $z_{\frac{1}{2}\alpha}$ may be used in place of $t_{N-r,\frac{1}{2}\alpha}$ where $(2\pi)^{-\frac{1}{2}} \int_{z_{\frac{1}{2}\alpha}}^{\infty} e^{-\frac{1}{2}x^2} dx = \frac{1}{2}\alpha$.

**Example 9 Finding a Confidence Interval for an Estimable Function** For the data of Examples 1–8 using the results of Examples 2 and 3 we have

$$\widehat{\alpha_1 - \alpha_2} = 46.725$$

and

$$v(\widehat{\alpha_1 - \alpha_2}) = \frac{5}{6}\sigma^2.$$

This holds true for estimators derived from generalized inverses $\mathbf{G}_1$ and $\mathbf{G}_2$.

From these results and using $\hat{\sigma}^2 = 241.666$ from Table 5.7, the symmetric $100(1 - \alpha)\%$ confidence interval on $\alpha_1 - \alpha_2$ is, from (47)

$$46.725 \pm \sqrt{241.666}t_{6-3,\frac{1}{2}\alpha}\sqrt{\tfrac{5}{6}}.$$
$$46.725 \pm \sqrt{241.666}(3.18245)\sqrt{\tfrac{5}{6}}$$
$$46.725 \pm 45.1626$$
$$(1.5624, 91.8876)$$

□

**d. What Functions Are Estimable?**

Whenever $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ for some $\mathbf{t}$ or $\mathbf{q}' = \mathbf{d}'\mathbf{U}'$ for some $\mathbf{d}$, then $\mathbf{q}'\mathbf{b}$ is estimable and has variance $\mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2$. We now consider some special cases.

Any linear function of $\mathbf{Xb}$ is estimable. Thus, for any vector $\mathbf{m}'$, say $\mathbf{m}'\mathbf{Xb}$, is estimable. Its b.l.u.e. is

$$\widehat{\mathbf{m}'\mathbf{Xb}} = \mathbf{m}'\mathbf{X}\mathbf{b}^\circ = \mathbf{m}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} \tag{48a}$$

with variance

$$v(\widehat{\mathbf{m}'\mathbf{Xb}}) = \mathbf{m}'\mathbf{X}'\mathbf{G}\mathbf{X}'\mathbf{m}\sigma^2. \tag{48b}$$

Any linear function of $\mathbf{X}'\mathbf{Xb}$ is also estimable because it is a linear function of $\mathbf{Xb}$, $\mathbf{s}'\mathbf{X}'\mathbf{Xb}$, say. Replacing $\mathbf{m}'$ in (48) by $\mathbf{s}'\ \mathbf{X}$ gives

$$\widehat{\mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{b}} = \mathbf{s}'\mathbf{X}'\mathbf{y} \tag{49a}$$

with variance

$$v(\widehat{\mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{b}}) = \mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{s}\sigma^2. \tag{49b}$$

Notice that $\mathbf{X}'\mathbf{Xb}$ is the same as the left-hand side of the normal equations with $\mathbf{b}^\circ$ replaced by $\mathbf{b}$. In addition, the b.l.u.e. of $\mathbf{s}'\mathbf{X}'\mathbf{Xb}$ is $\mathbf{s}'\mathbf{X}'\mathbf{y}$ where $\mathbf{X}'\mathbf{y}$ is the right-hand side of the normal equations. Based on these observations, we might in this sense, say that the b.l.u.e. of any linear function of the left-hand sides of the normal equations is the same function of the right-hand sides.

Linear functions of $E(\mathbf{b}^\circ)$ are also estimable, because $\mathbf{u}'E(\mathbf{b}^\circ) = \mathbf{u}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b}$. Using $\mathbf{u}'\mathbf{G}$ in place of $\mathbf{s}'$ in (49) shows that

$$\widehat{\mathbf{u}'E(\mathbf{b}^\circ)} = \mathbf{u}'\mathbf{G}\mathbf{X}'\mathbf{y} = \mathbf{u}'\mathbf{b}^\circ \tag{50a}$$

and

$$v[\widehat{\mathbf{u}'E(\mathbf{b}^\circ)}] = v(\mathbf{u}'\mathbf{b}^\circ) = \mathbf{u}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{u}\sigma^2 \tag{50b}$$

**TABLE 5.8 Estimable Functions and Their b.l.u.e.'s**

| Estimable Function | | | |
|---|---|---|---|
| Description | Function | b.l.u.e. | Variance of b.l.u.e. |
| General case: $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ | $\mathbf{q}'\mathbf{b}$ | $\mathbf{q}'\mathbf{b}^\circ$ | $\mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2$ |
| Linear function of $\mathbf{Xb}$ ($\mathbf{m}'$ arbitrary) | $\mathbf{m}'\mathbf{Xb}$ | $\mathbf{m}'\mathbf{Xb}^\circ$ | $\mathbf{m}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{m}$ |
| Linear function of $\mathbf{X}'\mathbf{Xb}$ ($\mathbf{s}'$ arbitrary) | $\mathbf{s}'\mathbf{X}'\mathbf{Xb}$ | $\mathbf{s}'\mathbf{X}'\mathbf{Xb}^\circ = \mathbf{s}'\mathbf{X}'\mathbf{y}$ | $\mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{s}\sigma^2$ |
| Linear function of $E(\mathbf{b}^\circ)$ ($\mathbf{u}'$ arbitrary) | $\mathbf{u}'E(\mathbf{b}^\circ)$ | $\mathbf{u}'\mathbf{b}^\circ$ | $\mathbf{u}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{u}\sigma^2$ |
| Vector $\mathbf{Hb}$ having $\mathbf{b}^\circ$ as b.l.u.e. | $\mathbf{Hb}$ | $\mathbf{b}^\circ$ | $\text{var}(\mathbf{b}^\circ) = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\sigma^2$ |

from (9). A special case of this result is when $\mathbf{u}'$ takes in turn the values of the rows of $\mathbf{I}$. In this instance, $\mathbf{b}^\circ$ is the b.l.u.e. of $\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b}$. These results are summarized in Table 5.8.

In view of the discussion of the $F$-statistics $F(R)$ and $F(R_{\text{m}})$ in Section 3, it is worth emphasizing two vectors that are not estimable, namely $\mathbf{b}$ and its sub-vector $\mathbf{b}_2$. They are not estimable because no value of $\mathbf{t}'$ where $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ can be found where $\mathbf{q}'\mathbf{b}$ reduces to an element of $\mathbf{b}$. Thus, no individual element of $\mathbf{b}$ is estimable. Therefore, $\mathbf{b}$ nor $\mathbf{b}$ is estimable.

### e. Linearly Independent Estimable Functions

From Table 5.8, it is evident that there are infinitely many estimable functions. How many linearly independent (LIN) estimable functions are there? The answer is that there are $r$ linearly independent estimable functions where $r$ is the rank of $\mathbf{X}$; that is, there are $r(\mathbf{X})$ LIN estimable functions.

Since $\mathbf{q}'\mathbf{b}$ with $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ is estimable for any $\mathbf{t}'$, let $\mathbf{T}'_{N\times N}$ be a matrix of full rank. Then, with $\mathbf{Q}' = \mathbf{T}'\mathbf{X}$, the functions $\mathbf{Q}'\mathbf{b}$ are $N$ estimable functions. However, $r(\mathbf{Q}) = r(\mathbf{X})$. Therefore, there are only $r(\mathbf{X})$ LIN rows in $\mathbf{Q}'$ and hence only $r(\mathbf{X})$ LIN terms in $\mathbf{Q}'\mathbf{b}$; that is, only $r(\mathbf{X})$ LIN estimable functions. Thus, any set of estimable functions cannot contain more than $r$ LIN such functions.

### f. Testing for Estimability

A given function $\mathbf{q}'\mathbf{b}$ is estimable if some vector $\mathbf{t}'$ can be found such that $\mathbf{t}'\mathbf{X} = \mathbf{q}'$. However, for $\mathbf{q}'$ known, derivation of a $\mathbf{t}'$ satisfying $\mathbf{t}'\mathbf{X} = \mathbf{q}'$ may not always be easy especially if $\mathbf{X}$ has large dimensions. Instead of deriving $\mathbf{t}'$, it can be determined whether $\mathbf{q}'\mathbf{b}$ is estimable by seeing whether $\mathbf{q}'$ satisfies the equation $\mathbf{q}'\mathbf{H} = \mathbf{q}'$, where $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$. We restate this as Theorem 2 below.

**Theorem 2** The linear function $\mathbf{q}'\mathbf{b}$ is estimable if and only if $\mathbf{q}'\mathbf{H} = \mathbf{q}'$.

*Proof.* If $\mathbf{q}'\mathbf{b}$ is estimable for some $\mathbf{t}'$, $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ so $\mathbf{q}'\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{H} = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{t}'\mathbf{X} = \mathbf{q}'$.

On the other hand, if $\mathbf{q}'\mathbf{H} = \mathbf{q}'$ then $\mathbf{q}' = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{t}'\mathbf{X}$ for $\mathbf{t}' = \mathbf{G}\mathbf{X}'$ so that $\mathbf{q}'\mathbf{b}$ is estimable. ■

Whether or not $\mathbf{q}'\mathbf{b}$ is estimable is easily established using Theorem 2. If $\mathbf{q}'\mathbf{b}$ is estimable, $\mathbf{q}'$ satisfies the equation $\mathbf{q}'\mathbf{H} = \mathbf{q}'$. Otherwise, it is not. Thus, we have a direct procedure for testing the estimability of $\mathbf{q}'\mathbf{b}$. Simply ascertain whether $\mathbf{q}'\mathbf{H} = \mathbf{q}'$. When $\mathbf{q}'\mathbf{H}$ does equal $\mathbf{q}'$, not only is $\mathbf{q}'\mathbf{b}$ estimable but from the last line of Table 5.8, the b.l.u.e. of $\mathbf{q}'\mathbf{b} = \mathbf{q}'\mathbf{H}\mathbf{b}$ is $\mathbf{q}'\mathbf{b}^\circ$. This corresponds to the invariance property of $\mathbf{q}'\mathbf{b}^\circ$ for $\mathbf{q}'\mathbf{H} = \mathbf{q}'$ derived in Theorem 6 of Chapter 1.

In developing the test, the generalized inverse $\mathbf{G}$ is completely arbitrary. An interesting condition can be obtained for estimability that uses the matrices from the singular value decomposition of $\mathbf{X}$. Suppose that $\mathbf{G}$ is the Moore–Penrose inverse. Then,

$$\mathbf{H} = (\mathbf{X}'\mathbf{X})^+\mathbf{X}'\mathbf{X} = \mathbf{U}\Lambda^{-1}\mathbf{U}'\mathbf{U}\Lambda\mathbf{U}' = \mathbf{U}\mathbf{U}'$$

and the condition $\mathbf{q}'\mathbf{H} = \mathbf{q}'$ reduces to

$$\mathbf{q}'\mathbf{U}\mathbf{U}' = \mathbf{q}' \tag{51a}$$

or

$$\mathbf{q}'\mathbf{V}\mathbf{V}' = \mathbf{0} \tag{51b}$$

since $\mathbf{U}\mathbf{U}' + \mathbf{V}\mathbf{V}' = \mathbf{I}$.

Thus, one way to determine if a linear combination is estimable is to obtain the eigenvector of 0 for $\mathbf{X}'\mathbf{X}$, normalize it to obtain $\mathbf{V}$ and then find $\mathbf{V}\mathbf{V}'$ and apply condition (51b). See Example 11 below.

**Example 10 Testing For Estimability** In Example 1, we had the generalized inverses

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_2 = \begin{bmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then,

$$\mathbf{H}_1 = \mathbf{G}_1\mathbf{X}'\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}_2 = \mathbf{G}_2\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We may use an $\mathbf{H}$ obtained from any generalized inverse. Consider the linear functions $-2\alpha_1 + \alpha_2 + \alpha_3$ and $\alpha_2 + \alpha_3$. Now,

$$[\,0 \quad -2 \quad 1 \quad 1\,]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = [\,0 \quad -2 \quad 1 \quad 1\,]$$

and

$$[\,0 \quad -2 \quad 1 \quad 1\,]\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [\,0 \quad -2 \quad 1 \quad 1\,].$$

However,

$$[\,0 \quad 0 \quad 1 \quad 1\,]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = [\,2 \quad 0 \quad 1 \quad 1\,]$$

and

$$[\,0 \quad 0 \quad 1 \quad 1\,]\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [\,0 \quad 0 \quad 1 \quad -1\,].$$

Hence, $-2\alpha_1 + \alpha_2 + \alpha_3$ is estimable but $\alpha_2 + \alpha_3$ is not.

From Example 2,

$$(\mathbf{b}_1^\circ)' = (\mathbf{G}_1\mathbf{X}'\mathbf{y}) = [\,0 \quad 47.51 \quad 0.785 \quad 0.19\,]$$

and

$$(\mathbf{b}_2^\circ)' = (\mathbf{G}_2\mathbf{X}'\mathbf{y}) = [\,0.19 \quad 47.32 \quad 0.595 \quad 0\,]$$

Thus, the b.l.u.e. of $-2\alpha_1 + \alpha_2 + \alpha_3$ is –94.045.

More generally, let $\mathbf{q}' = [\, q_1 \quad q_2 \quad q_3 \quad q_4 \,]$. Then,

$$\mathbf{q}'\mathbf{H}_1 = [\, q_1 \quad q_2 \quad q_3 \quad q_4 \,] \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = [\, q_2 + q_3 + q_4 \quad q_2 \quad q_3 \quad q_4 \,].$$

Then, $\mathbf{q}'\mathbf{b}$ is estimable if and only if $q_1 = q_2 + q_3 + q_4$. □

**Example 11 Using condition 51 to Determine Whether a Linear Function Is Estimable** Using $\mathbf{X}'\mathbf{X}$ from Example 1, we need to find the eigenvectors corresponding to the eigenvalue zero. If $\mathbf{X}'\mathbf{X}$ is $k \times k$, the multiplicity of the zero eigenvalues would be $k - r(\mathbf{X}'\mathbf{X})$. In this case, $r(\mathbf{X}'\mathbf{X}) = 3, k = 4$, so there is one zero eigenvalue. Let $\mathbf{v}' = [\, v_1 \;\; v_2 \;\; v_3 \;\; v_4 \,]$. Then,

$\mathbf{X}'\mathbf{X}\mathbf{v} = 0$ gives the system of equations

$$\begin{aligned} 6v_1 + 3v_2 + 2v_3 + v_4 &= 0 \\ 3v_1 + 3v_2 &= 0 \\ 2v_1 + 2v_3 &= 0 \\ v_1 + v_4 &= 0 \end{aligned}$$

A solution is $\mathbf{v}' = [\, 1 \;\; -1 \;\; -1 \;\; -1 \,]$. A normalized solution gives the desired matrix $\mathbf{V}' = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}$ so that

$$\mathbf{V}\mathbf{V}' = \frac{1}{4} \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix}.$$

Then,

$$\mathbf{q}'\mathbf{V}\mathbf{V}' = [\, q_1 \quad q_2 \quad q_3 \quad q_4 \,] \frac{1}{4} \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \frac{1}{4} \left[ q_1 - q_2 - q_3 - q_4 \quad -q_1 + q_2 + q_3 + q_4 \quad -q_1 + q_2 + q_3 + q_4 \quad -q_1 + q_2 + q_3 + q_4 \right]$$

and again the condition for estimability is $q_1 = q_2 + q_3 + q_4$. □

### g. General Expressions

In Table 5.8 and equations (48), $\mathbf{m'Xb}$ is estimable with b.l.u.e. $\mathbf{m'Xb}^\circ$ for any vector $\mathbf{m'}$ of order $N$. Thus, if we define $\mathbf{x}_j$ as the $j$th column of $\mathbf{X}$, then

$$\mathbf{X} = [\,\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_p\,]$$

and

$$\mathbf{m'Xb} = (\mathbf{m'x}_1)b_1 + (\mathbf{m'x}_2)b_2 + \cdots + (\mathbf{m'x}_p)b_p \tag{52}$$

with b.l.u.e.

$$\widehat{\mathbf{m'Xb}} = \mathbf{m'Xb}^\circ = (\mathbf{m'x}_1)b_1^\circ + (\mathbf{m'x}_2)b_2^\circ + \cdots + (\mathbf{m'x}_p)b_p^\circ. \tag{53}$$

For any values given to the $m_i$'s, the elements of $\mathbf{m}$, these same values, when used in (52), yield an estimable function, and when used in (52), they yield the b.l.u.e.'s of that estimable function.

Hence (52) and (53) constitute general expressions for an estimable function and its b.l.u.e.

Similar results hold for $\mathbf{s'X'Xs}$ of (49) where $\mathbf{s'}$ is any vector of order $p$, in distinction to $\mathbf{m'}$ of (52) and (53) which has order $N$. Defining $z_j$ as the $j$th column of $\mathbf{X'X}$,

$$\mathbf{s'X'Xb} = (\mathbf{s'}z_1)b_1 + (\mathbf{s'}z_2)b_2 + .... + (\mathbf{s'}z_p)b_p \tag{54}$$

with b.l.u.e.

$$\widehat{\mathbf{s'X'Xs}} = \mathbf{s'X'Xb}^\circ = (\mathbf{s'}z_1)b_1^\circ + (\mathbf{s'}z_2)b_2^\circ + \cdots + (\mathbf{s'}z_p)b_p^\circ. \tag{55}$$

The expressions in (54) and (55) hold for any elements in $\mathbf{s'}$ of order $p$ just as (52) and (53) hold for any elements of $\mathbf{m'}$ in order $N$.

From the last line of Table 5.8, we also have that $\mathbf{w'Hb}$ is estimable with b.l.u.e. $\mathbf{w'b}^\circ$. Thus, if

$$\mathbf{w'} = [\,w_1 \quad w_2 \quad \cdots \quad w_p\,]$$

and

$$\mathbf{H} = [\,\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_p\,],$$

then an estimable function is

$$\mathbf{w'Hb} = (\mathbf{w'h}_1)b_1 + (\mathbf{w'h}_2)b_2 + \cdots + (\mathbf{w'h}_p)b_p \tag{56}$$

with b.l.u.e.

$$\widehat{\mathbf{w}'\mathbf{H}\mathbf{b}} = \mathbf{w}'\mathbf{b}^{\circ} = w_1 b_1^{\circ} + w_2 b_2^{\circ} + \cdots + w_p b_p^{\circ}.. \tag{57}$$

Expressions (56) and (57) have advantages over (52) and (53) based on $\mathbf{m}'\mathbf{X}\mathbf{b}$ because of fewer arbitrary elements $p$ instead of $N$, and over (54) and (55) because of greater simplicity. This is evident in (57) which is just a linear combination of the elements of $\mathbf{b}^{\circ}$ where each element is multiplied by a single arbitrary $w$. Equation (55) often has a simple form too, because when $\mathbf{X}'\mathbf{X}$ is a design matrix, $\mathbf{H}$ often has $p - r$ null rows [$r = r(\mathbf{X})$], with its other $r$ rows having elements that are either 0, 1, or –1. The estimable function in (56) accordingly takes on a simple form and involves only $r$ elements of $\mathbf{w}$. Furthermore, in such cases, $\mathbf{b}^{\circ}$ can have only $r$ non-zero elements too, and so the b.l.u.e. in (57) then only involves $r$ terms.

We shall now establish that when $\mathbf{X}'\mathbf{X}$ is a design matrix, $\mathbf{H}$ can often be obtained as a matrix of 0's, 1's, and –1's. Suppose that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{X}_1'\mathbf{X}_1$ has full-row rank equal to $r(\mathbf{X})$, and $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Since $\mathbf{X} = [\,\mathbf{X}_1 \;\; \mathbf{X}_2\,]$, where $\mathbf{X}_1$ has full-column rank, $\mathbf{X}_2 = \mathbf{X}_1\mathbf{M}$ for some matrix $\mathbf{M}$, and because all elements of $\mathbf{X}$ are 0 or 1, those of $\mathbf{M}$ can often be 0, 1 or –1. Hence

$$\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{I} & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{M} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and so $p - r$ rows of $\mathbf{H}$ are null and the elements in the $r$ null rows are often 0, 1, or –1.

**Example 12 Numerical Illustration of Expressions for Estimable Functions** Recall from Examples 1–11 that the values of $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$, and $\mathbf{H}$ are

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{H}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

with

$$\mathbf{b} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \quad \text{and} \quad \mathbf{b}_1^\circ = \begin{bmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{bmatrix}.$$

With these values $\mathbf{m}'\mathbf{X}\mathbf{b}$ of (52) is

$$\begin{aligned} \mathbf{m}'\mathbf{X}\mathbf{b} = {} & (m_1 + m_2 + m_3 + m_4 + m_5 + m_6)\mu + (m_1 + m_2 + m_3)\alpha_1 \\ & + (m_4 + m_5)\alpha_2 + m_6\alpha_3 \end{aligned} \tag{58}$$

with b.l.u.e. from (53)

$$\widehat{\mathbf{m}'\mathbf{X}\mathbf{b}} = \mathbf{m}'\mathbf{X}\mathbf{b}_1^\circ = (m_1 + m_2 + m_3)47.51 + (m_4 + m_5)0.785 + m_6 0.19. \tag{59}$$

Thus for any values $m_1, m_2, \ldots, m_6$, (58) is an estimable function and (59) is its b.l.u.e.

Similarly from (54) and (55) using $\mathbf{X}'\mathbf{X}$,

$$\mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{b} = (6s_1 + 3s_2 + 2s_3 + s_4)\mu + 3(s_1 + s_2)\alpha_1 + 2(s_1 + s_3)\alpha_2 + (s_1 + s_4)\alpha_3 \tag{60}$$

is estimable with b.l.u.e.

$$\widehat{\mathbf{s}'\mathbf{X}'\mathbf{X}} = \mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{b}_1^\circ = 142.53(s_1 + s_2) + 1.57(s_1 + s_3) + 0.19(s_1 + s_4). \tag{61}$$

Expressions (60) and (61) hold true for any arbitrary values of the $s$'s. There are only $p = 4$ arbitrary $s$'s while there are $N = 6$ arbitrary $m$'s in (58) and (59). Expressions with fewer arbitrary values would seem preferable. Likewise, from (56) and (57), using $\mathbf{H}_1$, an estimable function is

$$\mathbf{w}_1'\mathbf{H}_1\mathbf{b} = (w_{11} + w_{13} + w_{14})\mu + w_{12}\alpha_1 + w_{13}\alpha_2 + w_{14}\alpha_3 \tag{62}$$

having b.l.u.e.

$$\widehat{\mathbf{w}_1'\mathbf{H}_1\mathbf{b}_1} = \mathbf{w}_1'\mathbf{b}_1^\circ = 47.51w_{12} + 0.785w_{13} + 0.19w_{14}. \tag{63}$$

For any values of $w_{12}, w_{13}$, and $w_{14}$, (62) is estimable and (63) is its b.l.u.e.

Note that in using (56) and (57), of which (62) and (63) are examples, the $\mathbf{H}$ used in $\mathbf{w}'\mathbf{H}\mathbf{b}$ in (56) must correspond to the $\mathbf{b}^\circ$ used in $\mathbf{w}'\mathbf{b}^\circ$ of (57). In (56), one cannot

use an $\mathbf{H}$ based in a generalized inverse that is different from the one used in deriving $\mathbf{b}_1^\circ = \mathbf{G}_1\mathbf{X}'\mathbf{y}$. This point is obvious, but important. Of course (56) and (57) apply for *any* $\mathbf{b}^\circ$ and its corresponding $\mathbf{H}$. Thus, for $\mathbf{b}_2^\circ$ and $\mathbf{H}_2$ equations (56) and (57) indicate that

$$\mathbf{w}_2'\mathbf{H}_2\mathbf{b} = w_{21}\mu + w_{22}\alpha_1 + w_{23}\alpha_2 + (w_{21} - w_{22} - w_{23})\alpha_3 \tag{64}$$

is estimable with b.l.u.e.

$$\widehat{\mathbf{w}_2'\mathbf{H}_2\mathbf{b}} = \mathbf{w}_2'\mathbf{b}_2^\circ = 0.19w_{21} + 47.32w_{22} + 0.595w_{23}. \tag{65}$$

The results in (65) hold for any values $w_{21}, w_{22}$, and $w_{23}$. Expressions (64) and (65) are not identical to (62) and (63). However, for different values of $w_{12}, w_{13}$, and $w_{14}$ and of $w_{21}, w_{22}$, and $w_{23}$ both pairs of expressions will generate the same set of estimable functions and their b.l.u.e.'s. For example, with $w_{12} = 0, w_{13} = 1$, and $w_{14} = 0$ equations (62) and (63) give $\mu + \alpha_2$ estimable with b.l.u.e. 0.785. Likewise, with $w_{21} = 1, w_{22} = 0$, and $w_{23} = 1$ equations (64) and (65) give $\mu + \alpha_2$ estimable with b.l.u.e.

$0.595(1) + 47.32(0) + 0.19(1) = 0.0785.$ □

## 5. THE GENERAL LINEAR HYPOTHESIS

In Section 6 of Chapter 3, we developed the theory for testing the general linear hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$ for the full-rank case. We shall now develop this theory for the non-full-rank case. In the non-full-rank case, we can test some hypotheses. Others, we cannot. We shall establish conditions for "testability" of a hypothesis.

### a. Testable Hypotheses

A testable hypothesis is one that can be expressed in terms of estimable functions. In Subsection d, we shall show that a hypothesis that is composed of non-estimable functions cannot be tested. It seems reasonable that a testable hypothesis should be made up of estimable functions because the results for the full-rank case suggest that $\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}$ will be part of the test statistic. If this is the case, $\mathbf{K}'\mathbf{b}^\circ$ will need to be invariant to $\mathbf{b}^\circ$. This can only happen if $\mathbf{K}'\mathbf{b}$ consists of estimable functions.

In light of the above considerations, a testable hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$ is taken as one, where $\mathbf{K}'\mathbf{b} \equiv \{\mathbf{k}_i'\mathbf{b}\}$ for $i = 1, 2, \ldots, m$ such that $\mathbf{k}_i'\mathbf{b}$ is estimable for all $i$. Hence $\mathbf{k}_i' = \mathbf{t}_i'\mathbf{X}$ for some $\mathbf{t}_i'$. As a result

$$\mathbf{K}' = \mathbf{T}'\mathbf{X} \tag{66}$$

for some matrix $(\mathbf{T})_{s\times N}$. Furthermore, any hypothesis is considered only in terms of its linearly independent components. Therefore $(\mathbf{K}')_{s\times p}$ is always of full-row rank.

Since $\mathbf{K}'\mathbf{b}$ is taken to be a set of estimable functions their b.l.u.e.'s are

$$\widehat{\mathbf{K}'\mathbf{b}} = \mathbf{K}'\mathbf{b}^{\circ} \tag{67a}$$

with expectation

$$E(\mathbf{K}'\mathbf{b}^{\circ}) = \mathbf{K}'\mathbf{b}. \tag{67b}$$

The b.l.u.e.'s have variance

$$\begin{aligned} \operatorname{var}(\widehat{\mathbf{K}'\mathbf{b}}) &= \mathbf{K}'\operatorname{var}(\mathbf{b}^{\circ})\mathbf{K} \\ &= \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{K}\sigma^2, \text{ from (9)} \\ &= \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{T}\sigma^2, \text{ from (66)} \\ &= \mathbf{K}'\mathbf{G}\mathbf{K}\sigma^2, \end{aligned} \tag{68}$$

making use of Theorem 10 of Chapter 1 and (66) again.

We shall now show that $\mathbf{K}'\mathbf{G}\mathbf{K}$ is non-singular. The functions $\mathbf{K}'\mathbf{b}$ are estimable. Thus $\mathbf{K}'$ can be represented not only as $\mathbf{T}'\mathbf{X}$ but also as $\mathbf{S}'\mathbf{X}'\mathbf{X}$ for some $\mathbf{S}'$ of full-row rank $m$. Then, with

$$\mathbf{K}' = \mathbf{S}'\mathbf{X}'\mathbf{X} \text{ of order } s \times p \quad \text{and} \quad r(\mathbf{K}') = s \leq r(\mathbf{X}),$$

since $\mathbf{K}'$ is of full-row rank $\mathbf{S}'$ and $\mathbf{S}'\mathbf{X}'$ have full-row rank $m$. Furthermore,

$$\mathbf{K}'\mathbf{G}\mathbf{K} = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{S} = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}.$$

Thus $r(\mathbf{K}'\mathbf{G}\mathbf{K}) = r(\mathbf{S}'\mathbf{X}') = s$, the order of $\mathbf{K}'\mathbf{G}\mathbf{K}$. Hence, $\mathbf{K}'\mathbf{G}\mathbf{K}$ is non-singular.

### b. Testing Testable Hypothesis

The test for the testable hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$ is developed just as in the full-rank case (Section 6a of Chapter 3). We assume that $\mathbf{e} \sim N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$. From Sections 3a and 3b, we have,

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}),$$

and

$$\mathbf{b}^{\circ} \sim N(\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{b}, \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}').$$

Furthermore, from (67) and (68),

$$\mathbf{K}\mathbf{b}^{\circ} - \mathbf{m} \sim N(\mathbf{K}'\mathbf{b} - \mathbf{m}, \mathbf{K}'\mathbf{G}\mathbf{K}).$$

Therefore using Theorem 5 of Chapter 2, the quadratic form,

$$Q = (\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}) \tag{69}$$

is such that

$$\frac{Q}{\sigma^2} \sim \chi^{2'}\left[s, \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{2\sigma^2}\right].$$

Furthermore,

$$Q = [\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'\mathbf{X}\mathbf{G}'\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{G}\mathbf{X}'[\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}],$$

with $(\mathbf{K}'\mathbf{K})^{-1}$ existing because $\mathbf{K}'$ has full-row rank, and

$$\begin{aligned}\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} &= \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}\\ &= \mathbf{T}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}\\ &= \mathbf{K}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} = \mathbf{m}.\end{aligned}$$

In addition,

$$\text{SSE} = [\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')[\mathbf{y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}],$$

because

$$\mathbf{X}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}') = \mathbf{0}. \tag{70}$$

Applying (70), we see that $Q$ and SSE are independent because the quadratic forms have null products. Therefore,

$$F(H) = \frac{Q/s}{\text{SSE}/(N-r)} \sim F'\left[s, N-r, \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{2\sigma^2}\right],$$

Under the null hypothesis $H\colon \mathbf{K}\mathbf{b} = \mathbf{m}$, the non-centrality parameter is zero. Thus, $F(H) \sim F_{s,N-r}$. Thus, $F(H)$ provides a test of the hypothesis $H\colon \mathbf{K}\mathbf{b} = \mathbf{m}$ with

$$F(H) = \frac{(\mathbf{K}'\mathbf{b} - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b} - \mathbf{m})}{s\hat{\sigma}^2} \tag{71}$$

with $s$ and $N - r$ degrees of freedom.

Suppose that we now seek a solution for $\mathbf{b}^\circ$ under the hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$. Denote it by $\mathbf{b}_H^\circ$. The solution will come from minimizing $(\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ)$

subject to $\mathbf{K}'\mathbf{b}_H^\circ = \mathbf{m}$. Using a Lagrange multiplier $2\theta'$ this leads exactly as in equation (117) of Chapter 3 to

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b}_H^\circ + \mathbf{K}'\theta &= \mathbf{X}'\mathbf{y} \\ \mathbf{K}'\mathbf{b}_H^\circ &= \mathbf{m}. \end{aligned} \tag{72}$$

From the first equation in (72), a solution is

$$\mathbf{b}_H^\circ = \mathbf{G}\mathbf{X}'\mathbf{y} - \mathbf{G}\mathbf{K}\theta. \tag{73}$$

Substitution of (73) into the second equation of (72) and following the derivation of equation (118) in Chapter 3 with the generalized inverse replacing the ordinary inverse we get

$$\mathbf{b}_H^\circ = \mathbf{b}^\circ - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}). \tag{74}$$

The error sum of squares after fitting this, denoted by $\mathrm{SSE}_H$, is

$$\begin{aligned} \mathrm{SSE}_H &= (\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ) \\ &= [\mathbf{y} - \mathbf{X}\mathbf{b}^\circ + \mathbf{X}(\mathbf{b}^\circ - \mathbf{b}_H^\circ)]'[\mathbf{y} - \mathbf{X}\mathbf{b}^\circ + \mathbf{X}(\mathbf{b}^\circ - \mathbf{b}_H^\circ)] \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ) + (\mathbf{b}^\circ - \mathbf{b}_H^\circ)'\mathbf{X}'\mathbf{X}(\mathbf{b}^\circ - \mathbf{b}_H^\circ). \end{aligned} \tag{75}$$

In deriving (75), the cross-product term vanishes because $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ) = \mathbf{0}$. Substituting from (74) for $\mathbf{b}^\circ - \mathbf{b}_H^\circ$, this gives

$$\mathrm{SSE}_H = \mathrm{SSE} + (\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}'\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}).$$

Now $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ and so

$$\mathbf{K}'\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{T}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{I}$$

and

$$\mathbf{K}'\mathbf{G}'\mathbf{K}' = \mathbf{T}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{T} = \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{T} = \mathbf{K}'\mathbf{G}\mathbf{K}.$$

Hence,

$$\begin{aligned} \mathrm{SSE}_H &= \mathrm{SSE} + (\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}) \\ &= \mathrm{SSE} + Q \end{aligned} \tag{76}$$

for $Q$ of (69).

**TABLE 5.9 Analysis of Variance for Testing the Hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$**

| Souce of variation | d.f. | Sum of Squares |
|---|---|---|
| Full model | $r$ | SSR $= \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$ |
| Hypothesis | $s$ | $Q = \mathbf{b}^{\circ\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^{\circ}$ |
| Reduced model | $r - s$ | SSR $- Q$ |
| Residual error | $N - r$ | SSE |
| Total | $N$ | SST $= \mathbf{y}'\mathbf{y}$ |

### c. The Hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$

For the non-full-rank model, we cannot apply the results in Section 5b to certain special cases as was done in Section 6c of Chapter 3 because (76) is limited to cases where $\mathbf{K}'\mathbf{b}$ is estimable. For example, we cannot test the hypotheses $H$: $\mathbf{b} = \mathbf{b}_0$ and $H$: $\mathbf{b}_q = \mathbf{0}$ because $\mathbf{b}$ and $\mathbf{b}_q$ are not estimable. Neither is $\tilde{\mathbf{b}}$. This is why as indicated in Section 3, tests based on $F(R)$ and $F(R_{\mathrm{m}})$ cannot be described as testing hypotheses of this nature. Nevertheless, as discussed in Section 2f(iii) of Chapter 6, the test based on $F(R_{\mathrm{m}})$ can sometimes be thought of as appearing equivalent to testing $\tilde{\mathbf{b}} = \mathbf{0}$.

One special case of the general hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is when $\mathbf{m} = \mathbf{0}$. Then $Q$ and $\mathbf{b}_H^{\circ}$ become

$$Q = \mathbf{b}^{\circ}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^{\circ} \quad \text{and} \quad \mathbf{b}_H^{\circ} = \mathbf{b}^{\circ} - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^{\circ}$$

with $Q =$ SSR – reduction in sum of squares due to fitting the reduced model.

Hence corresponding to Table 3.6 we have the analysis of variance shown in Table 5.9.

In Table 5.9, $r = r(\mathbf{K})$ and $s = r(\mathbf{K}')$, with $\mathbf{K}'$ having full row rank.
As before, we have three tests of hypothesis:

$$\frac{\mathrm{SSR}/r}{\mathrm{SSE}/(N-r)} \text{ tests the full model,}$$

$$\frac{Q/s}{\mathrm{SSE}/(N-r)} \text{ tests the hypothesis } H\text{: } \mathbf{K}'\mathbf{b} = \mathbf{0}$$

and under the null hypothesis,

$$\frac{(\mathrm{SSR} - Q)/(r-s)}{\mathrm{SSE}/(N-r)} \text{ tests the reduced model.}$$

The first and last of these tests are not to be construed as testing the fit of the models concerned but rather as testing their adequacy in terms of accounting for variation in the $y$ variable.

**TABLE 5.10 Analysis of Variance for Testing the Hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$ After Fitting the Mean**

| Souce of Variation | d.f. | Sum of Squares |
|---|---|---|
| Full Model | $r-1$ | $\text{SSR}_\text{m} = \text{SSR} - N\bar{y}^2$ |
| Hypothesis | $s$ | $Q = \mathbf{b}^{\circ\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ$ |
| Reduced Model (a.f.m.) | $r-s-1$ | $\text{SSR}_\text{m} - Q$ |
| Residual error | $N-r$ | SSE |
| Total (a.f.m.) | $N-1$ | $\text{SST}_\text{m} = \mathbf{y}'\mathbf{y} - N\bar{y}$ |

We can, of course, rewrite Table 5.9 to make it terms of "after fitting the mean" (a.f.m.). We do this by subtracting $N\bar{y}^2$ from SSR and SST to get $\text{SSR}_\text{m}$ and $\text{SST}_\text{m}$ as shown in Table 5.10. Again $r = r(\mathbf{K})$ and $s = r(\mathbf{K}')$, with $\mathbf{K}'$ having full row rank. The tests of hypotheses are then

$$\frac{\text{SSR}_\text{m}/(r-1)}{\text{SSE}/(N-r)} \text{ tests the full model (a.f.m.),}$$

$$\frac{Q/s}{\text{SSE}/(N-r)} \text{tests the hypothesis } H\colon \mathbf{K}'\mathbf{b} = \mathbf{0}$$

and under the null hypothesis,

$$\frac{(\text{SSR}_\text{m} - Q)/(r-s-1)}{\text{SSE}/(N-r)} \text{ tests the reduced model (a.f.m.).}$$

As was stated in the preceding paragraph, the first and the last of these tests relate to the adequacy of the model in explaining variation in the $y$ variable.

All of these results are analogous to those obtained for the full-rank model. In the non-full-rank case, we use $\mathbf{G}$ and $\mathbf{b}^\circ$ in place of $(\mathbf{X}'\mathbf{X})^{-1}$ and $\hat{\mathbf{b}}$ of the full-rank case. In fact, the full-rank model is just a special case of the non-full-rank model. When $\mathbf{X}'\mathbf{X}$ is non-singular, $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{b}^\circ = \hat{\mathbf{b}}$. All results for the full-rank model follow from those of the non-full-rank model.

### d. Non-testable Hypothesis

We noted earlier that a testable hypothesis is one composed of estimable functions. Our motivation was that we needed $\mathbf{K}\mathbf{b}^\circ$ to be invariant to $\mathbf{b}^\circ$ in order to be able to test $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$. What would happen if we tried to test a hypothesis that was not estimable? We illustrate with an example.

**Example 13 Attempt at Hypothesis Test With Non-Estimable Function** Consider the data from Examples 1–12. We shall attempt to test the non-estimable function $H\colon \alpha_2 = 0$ by calculating $Q$ in Table 5.10 for $\mathbf{G}_1$ and $\mathbf{G}_2$, and observing

that the answers are not the same. Using $\mathbf{b}_1^\circ, \mathbf{b}_2^\circ, \mathbf{G}_1$, and $\mathbf{G}_2$ as in Example 10, we have,

$$Q_1 = [\,0 \quad 47.51 \quad 0.785 \quad 0.19\,]\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\left[\left[\,0 \quad 0 \quad 1 \quad 0\,\right]\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & \frac{1}{3} & 0 & 0\\ 0 & 0 & \frac{1}{2} & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\right]^{-1}\cdot$$

$$[\,0 \quad 0 \quad 1 \quad 0\,]\begin{bmatrix}0\\47.51\\0.785\\0.19\end{bmatrix}$$

$$= 1.232$$

and

$$Q_2 = [\,0.19 \quad 47.32 \quad 0.595 \quad 0\,]\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\left[\left[\,0 \quad 0 \quad 1 \quad 0\,\right]\begin{bmatrix}1 & -1 & -1 & 0\\ -1 & \frac{4}{3} & 1 & 0\\ -1 & 1 & \frac{3}{2} & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\right]^{-1}\cdot$$

$$[\,0 \quad 0 \quad 1 \quad 0\,]\begin{bmatrix}0.19\\47.32\\0.595\\0\end{bmatrix}$$

$$= 0.236$$

Thus, the sum of squares due to the hypothesis and the reduced model would not be invariant to the choice of $\mathbf{b}^\circ$ and $\mathbf{G}$. Furthermore, for non-estimable hypotheses, the corresponding value of $\mathrm{SSE}_H$ is SSE and as a result, we cannot test the hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$. We now show this explicitly. □

The equations that result from minimizing $(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)$ subject to $\mathbf{K}'\mathbf{b}^\circ = \mathbf{m}$ are just as in (72),

$$\mathbf{X}'\mathbf{X}\mathbf{b}_H^\circ + \mathbf{K}\theta = \mathbf{X}'\mathbf{y} \quad \text{and} \quad \mathbf{K}'\mathbf{b}_H^\circ = \mathbf{m}, \tag{77}$$

where $2\theta'$ is a vector of Lagrange multipliers. Consider the equations

$$\mathbf{K}'(\mathbf{H} - \mathbf{I})\mathbf{z}_1 = \mathbf{m} - \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y} \tag{78}$$

in $\mathbf{z}_1\mathbf{b}$. As indicated in the proof of Theorem 4 of Chapter 1, $(\mathbf{H} - \mathbf{I})\mathbf{z}_1$ contains $p - r$ arbitrary elements. Since $\mathbf{K}'\mathbf{b}$ is not estimable, $\mathbf{K}' \neq \mathbf{T}'\mathbf{X}$ for any $\mathbf{T}'$. Thus, because $\mathbf{X} = \mathbf{X}'\mathbf{G}\mathbf{X}'\mathbf{X}$ (Theorem 10, Chapter 1), $\mathbf{K}' \neq (\mathbf{T}'\mathbf{X}\mathbf{G})\mathbf{X}'\mathbf{X}$, for any $\mathbf{T}'$. As a result,

the rows of $\mathbf{K}'$ are LIN of those of $\mathbf{X}'\mathbf{X}$. However, $\mathbf{X}'\mathbf{X}$ has order $p$ and rank $r$. Furthermore, the rows of $\mathbf{K}'$ have order $p$ and are to be LIN of each other. Therefore, if they are also to be LIN of the rows of $\mathbf{X}'\mathbf{X}$ there can be no more than $p - r$ of them. This means that $\mathbf{K}'$ has no more than $p - r$ rows. Hence (78) represents no more than $p - r$ equations in the $p - r$ unknowns of $(\mathbf{H} - \mathbf{I})\mathbf{z}_1$. Using it for $\mathbf{z}$ in

$$\mathbf{b}^\circ = \mathbf{G}\mathbf{X}'\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z} \tag{79}$$

to obtain

$$\mathbf{b}_H^\circ = \mathbf{G}\mathbf{X}'\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z}_1 \tag{80}$$

we find that $\theta = \mathbf{0}$ and that $\mathbf{b}_H^\circ$ of (80) satisfies (77). Consequently, because (80) is just a subset of the solutions (79) to $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$,

$$\text{SSE}_H = (\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}_H^\circ) = \text{SSE}$$

and so we cannot test the hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{m}$.

The sole difference between equations (72) and (76) is that in (72), $\mathbf{K}'\mathbf{b}$ is estimable while in (76), it is not. When solving (72), the estimability condition $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ for some $\mathbf{T}'$ leads to the solution (73). On the other hand, as shown in equations (79) and (80), the solution for $\mathbf{b}_H^\circ$ in (77) is also a solution to $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$. The lack of estimability of $\mathbf{K}'\mathbf{b}$ allows this. In contrast, in (72) where $\mathbf{K}'\mathbf{b}$ is estimable, $\mathbf{K}' = \mathbf{S}'\mathbf{X}'\mathbf{X}$ for some $\mathbf{S}'$. Then for $\mathbf{b}^\circ$ of (79), $\mathbf{K}'\mathbf{b}^\circ = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{b}_H^\circ = \mathbf{S}'\mathbf{X}'\mathbf{y}$ for all values of $\mathbf{z}$. Therefore, no value of $\mathbf{z}$ in (79) can be found such that $\mathbf{K}'\mathbf{b}^\circ = \mathbf{m}$. Thus, no value of $\mathbf{b}^\circ$ in (79) exists that satisfies (72).

Suppose we try to test a hypothesis that consists partly of estimable functions and partly of non-estimable functions? Assume $H\colon \mathbf{K}\mathbf{b} = \mathbf{m}$ can be written as

$$H: \begin{bmatrix} \mathbf{K}_1'\mathbf{b} \\ \mathbf{k}'\mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix} \tag{81}$$

where $\mathbf{K}_1'\mathbf{b}$ is estimable but $\mathbf{k}'\mathbf{b}$ is not. Then, using two Lagrange multipliers, the same development as above will lead to the conclusion that testing (81) is indistinguishable from just testing $H\colon \mathbf{K}_1'\mathbf{b} = \mathbf{m}_1$. Hence, in carrying out a test of hypothesis that consists partly of estimable functions and partly of non-estimable functions, all we are doing is testing the hypothesis made up of just estimable functions.

### e. Checking for Testability

The logic of deriving $Q = (\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})$ depends on $\mathbf{K}'\mathbf{b}$ being estimable. Nevertheless, when $\mathbf{K}'\mathbf{b}$ is not estimable, $Q$ can be calculated as long as $\mathbf{K}'\mathbf{G}\mathbf{K}$ is non-singular. This holds true, because estimability is a sufficient condition for the existence of $Q$, in particular, for the existence of $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$, but is not a

necessary condition. Hence, whenever $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ exists, $Q$ can be calculated even when $\mathbf{K}'\mathbf{b}$ is not estimable. Checking to see that $\mathbf{K}'\mathbf{b}$ is estimable is therefore essential before calculating $Q$ and $F(H)$. We have seen that there are a number of ways to do this including

1. ascertaining the existence of a matrix $\mathbf{T}'$ where $\mathbf{K}' = \mathbf{T}'\mathbf{X}$;
2. seeing if $\mathbf{K}'$ satisfies $\mathbf{K}' = \mathbf{K}'\mathbf{X}$;
3. ascertaining the existence of a matrix $\mathbf{C}'$, where $\mathbf{K}' = \mathbf{C}'\mathbf{U}'$, where $\mathbf{U}$ is the column orthogonal matrix in the singular value decomposition of $\mathbf{X}'\mathbf{X}$;
4. checking that $\mathbf{K}'$ satisfies either $\mathbf{K}'\mathbf{U}\mathbf{U}' = \mathbf{K}'$ or $\mathbf{K}'\mathbf{V}\mathbf{V}' = \mathbf{0}$.

Suppose, however, a researcher calculates $Q$ because he/she does not bother to check the estimability of $\mathbf{K}'\mathbf{b}$. If, in fact, $\mathbf{K}'\mathbf{b}$ is not estimable, what hypothesis, if any, is $F(H)$ testing? The answer is $H$: $\mathbf{K}'\mathbf{H}\mathbf{b} = \mathbf{m}$. We show this as follows. Since $H$: $\mathbf{K}'\mathbf{H}\mathbf{b} = \mathbf{m}$ is always testable, the value of $Q$, call it $Q_1$, is from (69),

$$Q_1 = (\mathbf{K}'\mathbf{H}\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{H}\mathbf{G}\mathbf{H}'\mathbf{K})^{-1}(\mathbf{K}'\mathbf{H}\mathbf{b}^\circ - \mathbf{m}). \tag{82}$$

In this expression $\mathbf{K}'\mathbf{H}\mathbf{b}^\circ = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{y} = \mathbf{K}'\mathbf{G}_1\mathbf{X}'\mathbf{y}$ because $\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{G}'\mathbf{X}'$ (Theorem 10 of Chapter 1) and where $\mathbf{G}_1 = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Therefore, $\mathbf{K}'\mathbf{H}\mathbf{b}^\circ = \mathbf{K}'\mathbf{G}_1\mathbf{X}'\mathbf{y} = \mathbf{K}'\mathbf{b}_1^\circ$, where $\mathbf{b}_1^\circ = \mathbf{G}_1\mathbf{X}'\mathbf{y}$ is a solution of $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$. Furthermore, $\mathbf{K}'\mathbf{H}\mathbf{G}\mathbf{H}'\mathbf{K} = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{K} = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K} = \mathbf{K}'\mathbf{G}_1\mathbf{K}$.

Thus, from (82) we obtain

$$Q_1 = (\mathbf{K}'\mathbf{b}_1^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}_1\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}_1^\circ - \mathbf{m}).$$

Thus, $Q_1$ is identical to the numerator sum of squares that would be calculated from (69) for testing the non-testable hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ using the solution $\mathbf{b}_1^\circ = \mathbf{G}_1\mathbf{X}'\mathbf{y}$. Hence, the calculations that might be made when trying to test the non-testable hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ are indistinguishable from those entailed in testing the testable hypothesis $\mathbf{K}'\mathbf{H}\mathbf{b} = \mathbf{m}$. In other words, if $F(H)$ of (71) is calculated for a hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ that is non-testable, the hypothesis actually being tested is $\mathbf{K}'\mathbf{H}\mathbf{b} = \mathbf{m}$. Example 14 below illustrates what has just been discussed.

**Example 14 Attempt at Testing the Non-estimable Function $\alpha_2 = 0$ from Example 13** According to what has been said already, an attempt to test $H$: $\alpha_2 = 0$ by calculating $Q_1$ in Example 13 would be equivalent to testing the hypothesis $\mu + \alpha_2 = 0$ because

$$[0 \quad 0 \quad 1 \quad 0]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = [1 \quad 0 \quad 1 \quad 0]$$

Then,

$$Q = \left[ [1 \quad 0 \quad 1 \quad 0] \begin{bmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{bmatrix} \right]' \left[ [1 \quad 0 \quad 1 \quad 0] \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right]^{-1} \cdot$$

$$\left[ [1 \quad 0 \quad 1 \quad 0] \begin{bmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{bmatrix} \right]$$

$$= 1.232$$

The generalized inverse $\mathbf{G}_1 = \mathbf{G}$ in this case because $\mathbf{G}$ is reflexive. □

### f. Some Examples of Testing Hypothesis

First, let us refresh ourselves on the results of some calculations in previous examples. We have

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b}_1^\circ = \begin{bmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{bmatrix}. \tag{83}$$

From (21)–(24)

$$\text{SSR} = 6772.87, \text{ SST} = 7497.87, \text{ and SSM} = 3469.93. \tag{84}$$

Thus,

$$\hat{\sigma}^2 = \frac{(7497.87 - 6772.87)}{3} = 241.667. \tag{85}$$

**Example 15 A Testable Hypothesis** Consider $H$: $\alpha_1 = \alpha_2 + 10$ or $\alpha_1 - \alpha_2 = 10$. It can be written $[0 \quad 1 \quad -1 \quad 0] \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = 10$.

Since

$$[0 \quad 1 \quad -1 \quad 0]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = [0 \quad 1 \quad -1 \quad 0],$$

the hypothesis is testable. We now calculate the $F$-statistic (71). We have that

$$\begin{aligned} \mathbf{k}'\mathbf{b}^{\circ} - \mathbf{m} &= 47.51 - .785 - 10 = 36.725, \\ \mathbf{k}'\mathbf{G}\mathbf{k} &= \tfrac{5}{6} \end{aligned} \tag{86}$$

and

$$F(H) = \frac{36.725(5/6)^{-1}36.725}{1(241.667)} = 4.6533$$

We fail to reject the hypothesis at $\alpha = .05$. □

We give another example of these computations for a testable hypothesis.

**Example 16 Another Testable Hypothesis** Consider $H : \mu + \alpha_1 = \mu + \alpha_2 = 50$. We may write this hypothesis as

$$\mathbf{K}'\mathbf{b} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 50 \\ 50 \end{bmatrix}.$$

We have that $\mathbf{K}'\mathbf{H}' = \mathbf{K}'$. Hence, the hypothesis is testable. Now

$$\mathbf{K}'\mathbf{b}^{\circ} - \mathbf{m} = \begin{bmatrix} 47.51 \\ 0.785 \end{bmatrix} - \begin{bmatrix} 50 \\ 50 \end{bmatrix} = \begin{bmatrix} -2.49 \\ -49.215 \end{bmatrix},$$

$$\mathbf{K}'\mathbf{G}\mathbf{K} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

and

$$F(H) = \frac{[\,-2.49 \quad -49.215\,]\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} -2.49 \\ -49.215 \end{bmatrix}}{2(241.667)} = 10.061.$$

We may write the same hypothesis as

$$\mathbf{K'b} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 50 \\ 0 \end{bmatrix}.$$

Then,

$$\mathbf{K'b}^\circ - \mathbf{m} = \begin{bmatrix} 47.51 \\ 46.725 \end{bmatrix} - \begin{bmatrix} 50 \\ 0 \end{bmatrix} = \begin{bmatrix} -2.49 \\ 46.725 \end{bmatrix}$$

and

$$\mathbf{K'GK} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{5}{6} \end{bmatrix}.$$

Hence,

$$F(H) = \frac{\begin{bmatrix} -2.49 & 46.725 \end{bmatrix}\begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} -2.49 \\ 46.725 \end{bmatrix}}{2(241.667)} = 10.061,$$

the same result. In this instance, we would reject $H$ at the .05 level of significance, the $p$-value being 0.047. □

**Example 17 A Hypothesis Test of the Form $\mathbf{K'b} = \mathbf{0}$ with Illustrations of Tables 5.9 and 5.10** We test the hypothesis $H$: $\alpha_1 = \alpha_2$ written as $[\,0 \;\; 1 \;\; -1 \;\; 0\,]\mathbf{b} = 0$. It is testable as seen in Example 15. As shown by (86), $\mathbf{k'Gk} = \frac{5}{6}$ and $\mathbf{k'b}^\circ - m = 36.725$. Then, $Q = 36.75^2(\frac{5}{6})^{-1} = 1620.68$. Table 5.9 then has the values from Table 5.11. If fitting the mean is to be taken into account as in Table 5.10 SSM is subtracted from SSR and SST to get $\text{SSR}_\text{m}$ and $\text{SST}_\text{m}$ as shown in Table 5.12. □

**TABLE 5.11 Example of Table 5.9**

| Source | d.f. | Sum of Squares |
|---|---|---|
| Full model | 3 | SSR = 6772.87 |
| Hypothesis | 1 | $Q = 1620.68$ |
| Reduced model | 2 | SSR – $Q = 5152.19$ |
| Residual error | 3 | SSE = 724.999 |
| Total | 6 | SST = 7497.86 |

### g. Independent and Orthogonal Contrasts

For a balanced linear model, linear combinations like $\alpha_1 + \alpha_2 - 2\alpha_3$ where the coefficients add up to zero are called contrasts. Given two contrasts, for example, the one above and $\alpha_1 - \alpha_2$ where the inner product of the coefficients are zero are said to be orthogonal. We shall now explore analogous notions for unbalanced data.

Recall that the numerator sum of squares for testing $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$

$$Q = \mathbf{b}^{\circ\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})\mathbf{K}'\mathbf{b}^{\circ}. \tag{87}$$

We shall see how to decompose $Q$ into a sum of squares involving individual orthogonal contrasts. Assume that $\mathbf{K}'\mathbf{b}$ is estimable. Then for some $\mathbf{S}'$, $\mathbf{K}' = \mathbf{S}'\mathbf{X}'\mathbf{X}$. With $\mathbf{b}^{\circ} = \mathbf{G}\mathbf{X}'\mathbf{y}$, using Theorem 10 of Chapter 1, we have that

$$\begin{aligned} Q &= \mathbf{y}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{X}\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\mathbf{y}. \end{aligned}$$

Furthermore $\mathbf{K}'$ has full-row rank $s$. When $s = r = r(\mathbf{X})$ it can be shown that $\mathbf{X}\mathbf{S} = \mathbf{X}_1\mathbf{P}\mathbf{X}'\mathbf{X}$, where $\mathbf{X}_1$, a sub-matrix of $\mathbf{X}$, is $N \times r$ of full-column rank, with $\mathbf{P}$ and $\mathbf{X}'_1\mathbf{X}_1$, both non-singular. This leads to $\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'$ being a generalized inverse of $\mathbf{X}'\mathbf{X}$ (see Exercise 11). Then,

$$Q = \mathbf{y}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} = \text{SSR when } s = r = r(\mathbf{X}). \tag{88}$$

**TABLE 5.12 Example of Table 5.10**

| Source | d.f. | Sum of Squares |
|---|---|---|
| Full model (a.f.m.) | 2 | $\text{SSR}_{\text{m}} = 3302.94$ |
| Hypothesis | 1 | $Q = 1620.68$ |
| Reduced model | 1 | $\text{SSR}_{\text{m}} - Q = 1682.68$ |
| Residual error | 3 | SSE = 724.999 |
| Total (a.f.m.) | 6 | $\text{SST}_{\text{m}} = 4027.94$ |

Now $r = r(\mathbf{X})$ is the maximum number of LIN estimable functions (see Section 4e). Hence (88) shows that the sum of squares SSR due to fitting the model $E(\mathbf{y}) = \mathbf{Xb}$ is exactly equivalent to the numerator sum of squares for testing the hypothesis $H\colon \mathbf{K}'\mathbf{b} = \mathbf{0}$ when $\mathbf{K}'\mathbf{b}$ represents the maximum number of LIN estimable functions, namely $r = r(\mathbf{X})$. This means that if $\mathbf{k}_j'$ is a row of $\mathbf{K}'$, then the numerator sum of squares for simultaneously testing $\mathbf{k}_i'\mathbf{b} = \mathbf{0}$ for $i = 1, 2, \ldots, r$ equals SSR. However, it does not necessarily mean that for testing the $r$ hypotheses $\mathbf{k}_i'\mathbf{b} = 0$, individually the sums add up to SSR. This will be true only in certain cases that we will now discuss.

Suppose that $\mathbf{k}_i'$ and $\mathbf{k}_j'$ are two rows of $\mathbf{K}'$. Then,

$$q_i = \mathbf{b}^{\circ\prime}\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{b}^\circ = \mathbf{y}'\mathbf{X}\mathbf{G}\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{y} \tag{89a}$$

and

$$q_j = \mathbf{b}^{\circ\prime}\mathbf{k}_j(\mathbf{k}_j'\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{b}^\circ = \mathbf{y}'\mathbf{X}\mathbf{G}\mathbf{k}_j(\mathbf{k}_j'\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{G}\mathbf{X}'\mathbf{y} \tag{89b}$$

are the numerator sums of squares for testing $\mathbf{k}_i'\mathbf{b} = 0$ and $\mathbf{k}_j'\mathbf{b} = 0$, respectively. Assume that $\mathbf{y} \sim N(\mathbf{Xb}, \sigma^2\mathbf{I})$. By Theorem 6 of Chapter 2, these sums of squares will be independent when $\mathbf{X}\mathbf{G}'\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{k}_j(\mathbf{k}_j'\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{G}\mathbf{X}' = 0$.

A necessary and sufficient condition for this is that $\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{k}_j = 0$.

Since $\mathbf{k}_j'\mathbf{b}$ is estimable, $\mathbf{k}_j' = \mathbf{t}'_j\mathbf{X}$, for some $\mathbf{t}'_j$. Thus, the condition becomes

$$\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}\mathbf{t}_j = \mathbf{k}_i'\mathbf{G}\mathbf{X}\mathbf{t}_j = \mathbf{k}_i'\mathbf{G}\mathbf{k}_j = 0. \tag{90}$$

Thus, (90) is a condition that makes $q_i$ and $q_j$ of (89), independent. Another important result follows from (90). Due to the independence of $q_i$ and $q_j$,

$$(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \operatorname{diag}\{(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i\} \quad \text{for} \quad i = 1, 2, \ldots, r.$$

Then (87) becomes

$$Q = \sum_{i=1}^{r} \mathbf{b}^{\circ\prime}\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{b}^\circ = \sum_{i=1}^{r} \frac{(\mathbf{k}_i'\mathbf{b}^\circ)^2}{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i} = \sum_{i=1}^{r} q_i. \tag{91}$$

By (45), condition (90) is also the condition that $\mathbf{k}_i'\mathbf{b}^\circ$ and $\mathbf{k}_j'\mathbf{b}^\circ$ are independent. Hence, $\mathbf{K}'\mathbf{b}$ consists of $r = r(\mathbf{X})$ LIN functions $\mathbf{k}_i'\mathbf{b}$ for $i = 1, 2, \ldots, r$. When, for $i = 1, 2, \ldots, r$,

$$\mathbf{k}_i' = \mathbf{k}_i'\mathbf{H}, \tag{92}$$

$$\mathbf{k}_i'\mathbf{G}\mathbf{k}_j = 0 \quad \text{for} \quad i \neq j \tag{93}$$

$$\text{the } \mathbf{k}_i' \text{ are LIN}, \tag{94}$$

then,

$$F(H) = \frac{Q}{r\hat{\sigma}^2} \quad \text{tests} \quad H\colon \mathbf{K}'\mathbf{b} = \mathbf{0},$$

and

$$F(H_i) = \frac{q_i}{\hat{\sigma}^2} \quad \text{tests} \quad H_i\colon \mathbf{k}_i'\mathbf{b} = 0,$$

and

$$Q = \text{SSR} = \sum_{i=1}^{r} q_i, \tag{95}$$

and the $q_i'$s are mutually independent with

$$q_i = \frac{(\mathbf{k}_i'\mathbf{b}^\circ)^2}{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i}.$$

Under their respective null hypotheses $F(H) \sim F_{r,N-r}$ and $F(H_i) \sim F_{1,N-r}$. Using the latter of these two $F$-statistics is equivalent to performing a $t$-test with $N - r$ degrees of freedom. The $t$-statistic that is used to test $H_i$ is

$$\sqrt{\frac{q_i}{\hat{\sigma}^2}} = \frac{\mathbf{k}_i'\mathbf{b}^\circ}{\sqrt{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i\hat{\sigma}^2}}.$$

For balanced data, these conditions lead to sets of values for the $\mathbf{k}_i'$ such that the $\mathbf{k}_i'\mathbf{b}$ are often called orthogonal contrasts. They are "orthogonal" because $\mathbf{G}$ is such that (93) reduces to $\mathbf{k}_i'\mathbf{k}_j = 0$. They are called "contrasts" because the $\mathbf{k}_i'\mathbf{b}$ can be expressed as sums of differences between the elements of $\mathbf{b}$. We retain the name "orthogonal contrasts" here for unbalanced data meaning orthogonal in the sense of (93). Examples are given below and in Chapter 6.

### h. Examples of Orthogonal Contrasts

First, let us consider an example for the balanced case.

**Example 18 Orthogonal Contrasts for a Balanced Model** Consider the linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ of the form

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_3 & \mathbf{1}_3 & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_3 & \mathbf{0} & \mathbf{1}_3 & \mathbf{0} \\ \mathbf{1}_3 & \mathbf{0} & \mathbf{0} & \mathbf{1}_3 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + e.$$

Then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 9 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}$$

has a generalized inverse

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}$$

and

$$\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Let $\mathbf{q}' = [\, q_1 \;\; q_2 \;\; q_3 \;\; q_4 \,]$. Then $\mathbf{q}'\mathbf{b}$ is estimable if $\mathbf{q}'\mathbf{H} = \mathbf{q}'$ or when $q_1 = q_2 + q_3 + q_4$. Examples of estimable functions include $\mu + \alpha_i, i = 1, 2, 3$ and $\alpha_i - \alpha_j, i \neq j$. Let $\mathbf{p}' = [\, p_1 \;\; p_2 \;\; p_3 \;\; p_4 \,]$ and let $\mathbf{p}'\mathbf{b}$ be an estimable function. Contrasts are differences or sums of differences like $\alpha_1 - \alpha_2$ and $(\alpha_1 - \alpha_3) + (\alpha_2 - \alpha_3) = \alpha_1 + \alpha_2 - 2\alpha_3$. The orthogonality condition (93) reduces to $\sum_{i=2}^{4} p_i q_i = 0$. The two contrasts mentioned above are clearly orthogonal. □

We now consider an example for the unbalanced case.

**Example 19 Orthogonal Contrasts for Unbalanced Data** For the $\mathbf{X}$ matrix considered in Examples 1–17, we have that $r(\mathbf{X}) = r = 3$. To illustrate $Q$ and SSR in (88), we consider the hypothesis $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$ for

$$\mathbf{K}' = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

The rows of $\mathbf{K}'$ are LIN. Furthermore, because $\mathbf{K}'\mathbf{H} = \mathbf{K}'$, the elements of $\mathbf{K}'\mathbf{b}$ are estimable. Using $\mathbf{b}^\circ$ and $\mathbf{G}$ of (83) from (87), the numerator sum of squares is

$$Q = [\,48.485 \quad 94.045 \quad 0.595\,] \left[ \frac{1}{6} \begin{bmatrix} 11 & -5 & -3 \\ -5 & 17 & 3 \\ -3 & 3 & 9 \end{bmatrix} \right]^{-1} \begin{bmatrix} 48.485 \\ 94.045 \\ 0.595 \end{bmatrix} = 6772.87 = \text{SSR}$$

in Table 5.11. Simultaneous testing of the hypotheses

$$\begin{aligned} H_1&: 3\mu + \alpha_1 + \alpha_2 + \alpha_3 = 0 \\ H_2&: 2\alpha_1 - \alpha_2 - \alpha_3 = 0 \\ H_3&: \alpha_2 - \alpha_3 = 0 \end{aligned}$$

uses a numerator sum of squares equal to SSR. However, adding the numerator sum of squares for testing these hypotheses individually does not give SSR as shown below.

| Hypothesis | Numerator Sum of Squares |
|---|---|
| $3\mu + \alpha_1 + \alpha_2 + \alpha_3 = 0$ | $\frac{48.485^2}{11/6} = 1282.25$ |
| $2\alpha_1 - \alpha_2 - \alpha_3 = 0$ | $\frac{94.045^2}{17/6} = 3121.57$ |
| $\alpha_2 - \alpha_3 = 0$ | $\frac{0.595^2}{9/6} = 0.24$ |
| Total | $4404.06 \neq 6772.87$ |

For balanced data, the individual hypotheses of $\mathbf{K}'\mathbf{b} = \mathbf{0}$, given above, would be considered orthogonal contrasts. This is not the case for unbalanced data because the b.l.u.e.'s of the estimable functions involved in the hypotheses are not distributed independently. Their covariance matrix does not have zero-off diagonal elements as seen below. We have that

$$\text{var}(\mathbf{K}'\mathbf{b}^\circ) = \mathbf{K}'\mathbf{G}\mathbf{K}\sigma^2 = \frac{1}{6} \begin{bmatrix} 11 & -5 & -3 \\ -5 & 17 & 3 \\ -3 & 3 & 9 \end{bmatrix} \sigma^2$$

For balanced data, $\mathbf{K}'\mathbf{G}\mathbf{K}$ would be diagonal giving rise to independence.

We shall derive a set of orthogonal contrasts that satisfy (93). To do this, we need to obtain $\mathbf{K}'$ so that its rows satisfy (92)–(94). Suppose that one contrast of interest

is $\alpha_1 - \alpha_3$. In order to find two other contrasts that are orthogonal to it, we take $\mathbf{K}'$ to have the form

$$\mathbf{K}' = \begin{bmatrix} a & b & c & d \\ 0 & 1 & 0 & -1 \\ 0 & f & g & h \end{bmatrix}$$

Using $\mathbf{H}$ in (83) in order to have $\mathbf{K}'\mathbf{H} = \mathbf{K}'$, the condition for estimability, (92) demands that

$$b + c + d = a \quad \text{and} \quad f + g + h = 0.$$

The conditions in (93) gives

$$\tfrac{1}{3}b - d = 0, \quad \tfrac{1}{3}f - h = 0, \quad \tfrac{1}{3}bf + \tfrac{1}{2}cg + dh = 0.$$

For any values of $d$ and $h$ solutions to these two sets of equations are

$$\tfrac{1}{6}a = \tfrac{1}{3}b = \tfrac{1}{2}c = d \quad \text{and} \quad \tfrac{1}{3}f = -\tfrac{1}{4}g = h.$$

For example, putting $d = 1$ and $h = 1$ gives

$$\mathbf{K}' = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 3 & -4 & 1 \end{bmatrix}$$

Then,

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} 144.29 \\ 47.32 \\ 139.58 \end{bmatrix} \quad \text{and} \quad \mathbf{K}'\mathbf{G}\mathbf{K} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & \frac{4}{3} & 0 \\ 0 & 0 & 12 \end{bmatrix}.$$

Notice that $\mathbf{K}'\mathbf{G}\mathbf{K}$ above has off-diagonal elements zero. Thus, $\mathbf{K}$ satisfies (93) and the contrasts are orthogonal. Furthermore, the rows of $\mathbf{K}'$ are LIN and

thus satisfy (94). We test the hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$ using (87). Calculating $Q$, we have

$$Q = [\,144.29 \quad 47.32 \quad 139.58\,]\begin{bmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{3}{4} & 0 \\ 0 & 0 & \frac{1}{12} \end{bmatrix}\begin{bmatrix} 144.29 \\ 47.32 \\ 139.58 \end{bmatrix}$$
$$= \frac{144.29^2}{6} + 47.32^2(\tfrac{3}{4}) + \frac{(139.58)^2}{12} = 3469.93 + 1679.39 + 1623.55$$
$$= 6772.87.$$

which is equal to SSR of Table 5.11.

From this development, we see that estimable and LIN contrasts

$$c_1 = 6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3$$
$$c_2 = \alpha_1 - \alpha_3$$
$$c_3 = 3\alpha_1 - 4\alpha_2 + \alpha_3$$

are orthogonal in the manner of (93). Furthermore, the numerator sums of squares for testing each of them add up to that for testing them simultaneously, namely SSR. This illustrates (95).

Notice that for testing $H$: $6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3 = 0$, the numerator sum of squares is $144.29^2/6 = 3469.93 = N\bar{y}^2 = \text{SSM}$. Furthermore, the sums of squares for the contrasts orthogonal to this, 1679.39 and 1623.55, sum to 3302.94, $\text{SSR}_\text{m}$, the sum of squares due to fitting the model correcting for the mean (see Table 5.12). In general, consider any contrast $\mathbf{k}'\mathbf{b}$ that is orthogonal to $6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3$. By (92) with $\mathbf{H}$ of (83), the form of $\mathbf{k}'$ must be

$$\mathbf{k}' = [\,k_2 + k_3 + k_4 \quad k_2 \quad k_3 \quad k_4\,].$$

The condition in (93) requires that $\mathbf{k}'$ must satisfy

$$\mathbf{k}'\mathbf{G}\begin{bmatrix} 6 \\ 3 \\ 2 \\ 1 \end{bmatrix} = [\,k_2 + k_3 + k_4 \quad k_2 \quad k_3 \quad k_4\,]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 6 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$
$$= k_2 + k_3 + k_4 = 0.$$

Thus $\mathbf{k}' = [\,0 \;\; k_2 \;\; k_3 \;\; k_4\,]$ with $k_2 + k_3 + k_4 = 0$. Thus, any contrast $\mathbf{k}'\mathbf{b}$ with $k_2 + k_3 + k_4 = 0$ that satisfies (92) and (93) is orthogonal in the manner of (93) to $6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3$ and, because the first term is zero, does not involve $\mu$.

One such contrast is $2\alpha_1 - \alpha_2 - \alpha_3$. Any $r - 1$ such contrasts that are orthogonal to each other will have numerator sums of squares that sum to $\mathrm{SSR_m}$. For example, if

$$\mathbf{K}' = \begin{bmatrix} 0 & 2 & -1 & -1 \\ 0 & a & b & c \end{bmatrix}$$

$\mathbf{K}'\mathbf{b}$ will be a pair of orthogonal contrasts, orthogonal to each other and to $6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3$, if $a + b + c = 0$ and

$$[\,0 \quad 2 \quad -1 \quad -1\,]\mathbf{G}\begin{bmatrix} 0 \\ a \\ b \\ c \end{bmatrix} = \tfrac{2}{3}a - \tfrac{1}{2}b - c = 0.$$

One solution to this system of equations is $a = 3$, $b = -10$, $c = 7$. For this solution,

$$\mathbf{K}' = \begin{bmatrix} 0 & 2 & -1 & -1 \\ 0 & 3 & -10 & 7 \end{bmatrix}.$$

Then,

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} 94.045 \\ 136.01 \end{bmatrix} \quad \text{and} \quad \mathbf{K}'\mathbf{G}\mathbf{K} = \begin{bmatrix} \frac{17}{6} & 0 \\ 0 & 102 \end{bmatrix}.$$

Then in (87), we have,

$$Q = \frac{94.045^2}{17/6} + \frac{136.01^2}{102} = 3302.93 = \mathrm{SSR_m}$$

of Table 5.12 with a slight round off error. □

The above examples illustrate the several ways in which (92)–(95) can be used for establishing independent and orthogonal contrasts for unbalanced data and testing hypotheses about them. We shall give more examples in Chapter 6.

## 6. RESTRICTED MODELS

We have observed that sometimes a linear model may include restrictions on the parameter vectors. Such restrictions are quite different from the "usual constraints". The "usual constraints" are frequently introduced for the sole purpose of obtaining a solution to the normal equations. We will discuss this in Section 7. In contrast, we shall consider the restrictions that we present here to be an integral part of the model.

As such, these restrictions must be taken into account in the estimation and testing processes.

So far, the discussion has been in terms of models whose parameters have been very loosely defined. Indeed, no formal definitions have been made. When writing the equation of the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, we simply described $\mathbf{b}$ as being the vector of parameters of the model and left it at that. Thus, in the examples, $\mu$ is described simply as a general mean and $\alpha_1, \alpha_2,$ and $\alpha_3$ as the effect on yield arising from three different plant varieties. We imply no further definition. Sometimes, however, more explicit definitions inherent in the model result in relationships (or restrictions) existing among the parameters of the model. They are considered part and parcel of the model. For example, the situation may be such that the parameters of the model satisfy $\alpha_1 + \alpha_2 + \alpha_3 = 0$. We take this not as a hypothesis but as a fact without question. We will call these kinds of relationships that exist as an integral part of the model *restrictions on the model*. Their origin and concept are not the same as those of relationships that we sometimes impose on the normal equations in order to simplify, obtaining their solution. Those relationships will be called *constraints on the solutions*. We shall discuss these in Section 7. Here we concern ourselves with an aspect of the model. It includes relationships among its parameters. One simple example might be a model involving three angles of a triangle.
Another might involve the total weight and its components, such as fat, bone, muscle, and lean meat in a dressed beef carcass.

The models already discussed, those that contain no restrictions of the kind just referred to, will be referred to as *unrestricted models*. Models that do include restrictions of this nature will be called *restricted models*. The question then arises as to how the estimation and testing hypothesis processes developed for unrestricted models apply to restricted models. In general, we consider the set of restrictions

$$\mathbf{P}'\mathbf{b} = \boldsymbol{\delta} \tag{96}$$

as part of the models, where $\mathbf{P}'$ has row rank $q$. The restricted model is then $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ subject to the restriction $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$. Fitting this restricted model leads, just as in (72), to

$$\mathbf{X}'\mathbf{X}\mathbf{b}_r^\circ + \mathbf{P}\theta = \mathbf{X}'\mathbf{y} \tag{97a}$$

and

$$\mathbf{P}'\mathbf{b}_r^\circ = \boldsymbol{\delta}. \tag{97b}$$

Again $2\theta$ is a vector of Lagrange multipliers. The subscript $r$ on $\mathbf{b}_r^\circ$ denotes that $\mathbf{b}_r^\circ$ is a solution to the normal equations of the restricted model. To solve (97), we must make a distinction as to whether in the unrestricted model $\mathbf{P}'\mathbf{b}$ is estimable or not estimable because the solution is not the same in the two cases. We first consider the case where $\mathbf{P}'\mathbf{b}$ is estimable.

### a. Restrictions Involving Estimable Functions

When $\mathbf{P}'\mathbf{b}$ is estimable, we have by analogy with (73) that a solution to (97) is

$$\mathbf{b}_r^\circ = \mathbf{b}^\circ - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{b}^\circ - \boldsymbol{\delta}). \tag{98}$$

Its expected value is

$$E(\mathbf{b}_r^\circ) = \mathbf{Hb} - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{Hb} - \boldsymbol{\delta}) = \mathbf{Hb}. \tag{99}$$

To obtain (99), we use $E(\mathbf{b}^\circ) = \mathbf{Hb}$ of (8), $\mathbf{P}'\mathbf{H} = \mathbf{P}'$ because $\mathbf{P}'\mathbf{b}$ is estimable and (96).

After some simplification (see Exercise 14), the variance of $\mathbf{b}_r^\circ$ is

$$\operatorname{var}(\mathbf{b}_r^\circ) = \operatorname{var}\{[\mathbf{I} - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}']\mathbf{b}^\circ\} = \mathbf{G}[\mathbf{X}'\mathbf{X} - \mathbf{P}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}']\mathbf{G}'\sigma^2. \tag{100}$$

The error sum of squares after fitting this restricted model is

$$\mathrm{SSE}_r = (\mathbf{y} - \mathbf{Xb}_r^\circ)'(\mathbf{y} - \mathbf{Xb}_r^\circ).$$

From (75) and (76), we see that

$$\mathrm{SSE}_r = \mathrm{SSE} + (\mathbf{P}'\mathbf{b}^\circ - \boldsymbol{\delta})'(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{b}^\circ - \boldsymbol{\delta}) \tag{101a}$$

with

$$E(\mathrm{SSE}_r) = (N - r)\sigma^2 + E\mathbf{b}^{\circ\prime}\mathbf{P}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}'\mathbf{b}^\circ - \boldsymbol{\delta}'(\mathbf{P}'\mathbf{GP})^{-1}\boldsymbol{\delta}. \tag{101b}$$

We apply Theorem 4 of Chapter 2 to the middle term of (101b). Using (8) and (86) and (96) again, (101b) reduces to

$$E(\mathrm{SSE}_r) = (N - r + q)\sigma^2.$$

Hence, in the restricted model, an unbiased estimator of the error variance is

$$\hat{\sigma}_r^2 = \frac{\mathrm{SSE}_r}{N - r + q}. \tag{102}$$

(There should be no confusion over the letter $r$ used as the rank of $\mathbf{X}$ and as a subscript to denote restricted.)

Observe that $\mathbf{b}_r^\circ$ and $\mathrm{SSE_r}$ of (98) and (101) are not the same as $\mathbf{b}^\circ$ and SSE. This indicates that estimable restrictions on the parameters of the model affect the estimation process. However, this does not affect the estimability of any function that is estimable in the unrestricted model. Thus, if $\mathbf{k}'\mathbf{b}$ is estimable in the unrestricted model, it is still estimable in the restricted model. The condition for estimability, that

is for some $\mathbf{t}'$, $E(\mathbf{t}'\mathbf{y}) = \mathbf{k}'\mathbf{b}$, remains unaltered. However, although the function is still estimable, it is a function of the parameters and therefore subject to the restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$. These may change the form of $\mathbf{k}'\mathbf{b}$. Consider, for example the function $\mathbf{k}'\mathbf{b} = \mu + \frac{1}{2}(\alpha_1 + \alpha_2)$. It is estimable. However, in a restricted model having the restriction $\alpha_1 - \alpha_2 = 0$, $\mathbf{k}'\mathbf{b}$ becomes $\mu + \alpha_1$ or equivalently $\mu + \alpha_2$.

Given the restriction $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$, in general, the estimable function is changed to $\mathbf{k}'\mathbf{b} + \boldsymbol{\lambda}'(\mathbf{P}'\mathbf{b} - \delta)$. In order that this just be a function of the $b$'s, $\boldsymbol{\lambda}'$ must be such that $\boldsymbol{\lambda}'\boldsymbol{\delta} = 0$. (When $\boldsymbol{\delta} = \mathbf{0}, \boldsymbol{\lambda}'$ can be any vector.) Then $\mathbf{k}'\mathbf{b}$ becomes $\mathbf{k}'\mathbf{b} + \boldsymbol{\lambda}'\mathbf{P}'\mathbf{b} = (\mathbf{k}' + \boldsymbol{\lambda}'\mathbf{P}')\mathbf{b}$. Of course, this is estimable for the unrestricted model because both $\mathbf{k}'\mathbf{b}$ and $\mathbf{P}'\mathbf{b}$ are.

In the restricted model, the hypothesis $H$: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ can be considered only if it is consistent with $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$. For example, if $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$ is $\alpha_1 - \alpha_2 = 0$, one cannot consider the hypothesis $\alpha_1 - \alpha_2 = 4$. Within this limitation of consistency, the hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is tested in the restricted model by considering the unrestricted model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ subject to both the restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$ and the testable hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$. The restricted model reduced by the hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ can be called the *reduced restricted model*. On writing

$$\mathbf{Q}' = \begin{bmatrix} \mathbf{P}' \\ \mathbf{K}' \end{bmatrix} \quad \text{and} \quad \boldsymbol{\ell} = \begin{bmatrix} \boldsymbol{\delta} \\ \mathbf{m} \end{bmatrix}$$

we minimize $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ subject to $\mathbf{Q}'\mathbf{b} = \boldsymbol{\ell}$.
Since both $\mathbf{P}'$ and $\mathbf{K}'$ have full-row rank and their rows are mutually LIN, $\mathbf{Q}'$ has full-row rank and $\mathbf{Q}'\mathbf{b}$ is estimable. The minimization leads to the solution $\mathbf{b}^\circ_{r,H}$. From (74) this is

$$\mathbf{b}^\circ_{r,H} = \mathbf{b}^\circ - \mathbf{G}\mathbf{Q}(\mathbf{Q}'\mathbf{G}\mathbf{Q})^{-1}(\mathbf{Q}'\mathbf{b}^\circ - \boldsymbol{\ell}).$$

The corresponding residual sum of squares is

$$\mathrm{SSE}_{r,H} = \mathrm{SSE} + (\mathbf{Q}'\mathbf{b}^\circ - \boldsymbol{\ell})'(\mathbf{Q}'\mathbf{G}\mathbf{Q})^{-1}(\mathbf{Q}'\mathbf{b}^\circ - \boldsymbol{\ell}).$$

The test of hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is based on

$$F(H_r) = \frac{(\mathrm{SSE}_{r,H} - \mathrm{SSE}_r)}{s\hat{\sigma}^2_r} \tag{103}$$

where $\hat{\sigma}^2_r = \mathrm{SSE}_r/(N - r + q)$ as in (102).

Recall that a function that is estimable in an unrestricted model is estimable in the restricted model. Likewise, a hypothesis that is testable in an unrestricted model is also testable in the restricted model. The form of the hypothesis may be changed as a result of the restrictions. Nevertheless, the modified form of the hypothesis will be tested under both the restricted and the unrestricted model.

**Example 20 Hypotheses that are Testable under Restricted and Unrestricted Models** The hypothesis $H$: $\mu + \frac{1}{2}(\alpha_1 + \alpha_2) = 20$ is testable in the unrestricted model. In a restricted model having $\alpha_1 - \alpha_2 = 4$ as a restriction, the hypothesis is modified to be

$H$: $\mu + \alpha_2 = 18$ or $H$: $\mu + \alpha_1 = 22$. These are testable in the restricted model. They are also testable in the unrestricted model. □

In general, if $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is testable in the unrestricted model, then, for any matrix $\mathbf{L}_{s\times q}$, $(\mathbf{K}' + \mathbf{L}\mathbf{P}')\mathbf{b} = \mathbf{m} + \mathbf{L}\delta$ will be testable in the restricted model. It will also be testable in the unrestricted model.

### b. Restrictions Involving Non-estimable Functions

When the restrictions are $\mathbf{P}'\mathbf{b} = \delta$ and $\mathbf{P}'\mathbf{b}$ is not estimable, the solutions to (97) are similar to (80),

$$\mathbf{b}_r^\circ = \mathbf{b}^\circ + (\mathbf{H} - \mathbf{I})\mathbf{z}_1 \tag{104}$$

where, following (78), $\mathbf{z}_1$ satisfies

$$\mathbf{P}'(\mathbf{H} - \mathbf{I})\mathbf{z}_1 = \delta - \mathbf{P}'\mathbf{G}\mathbf{X}'\mathbf{y}. \tag{105}$$

Hence, $\mathbf{b}^\circ$ is just one of the solutions to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$. Therefore, in this case, $\text{SSE}_r = \text{SSE}$. The restrictions do not affect the residual sum of squares.

Just as before, the inclusion of restrictions in the model does not alter the estimability of a function that is estimable in the unrestricted model. It is still estimable in the restricted model. However, it will be amended because of the restrictions. Since the restrictions do not involve estimable functions, the amended form of an estimable function may be such that even though it is estimable in the restricted model, it is not estimable in the unrestricted model. Consider the model used for Examples 1–17. The function $\mu + \frac{1}{2}(\alpha_1 + \alpha_2)$ is estimable in the unrestricted model. However, for a restricted model that includes the restriction $\alpha_1 = 0$, we amend the function $\mu + \frac{1}{2}(\alpha_1 + \alpha_2)$ to be $\mu + \frac{1}{2}\alpha_2$. This amended function, although estimable in the restricted model, is not estimable in the unrestricted model.

Thus, functions that are not estimable in unrestricted models may be estimable in restricted models. In general, if $\mathbf{k}'\mathbf{b}$ is estimable in the unrestricted model, then $\mathbf{k}'\mathbf{b} + \boldsymbol{\lambda}'(\mathbf{P}'\mathbf{b} - \delta)$ is estimable in the restricted model provided that either $\delta = \mathbf{0}$ or $\boldsymbol{\lambda}'$ is such that $\boldsymbol{\lambda}'\delta = 0$. Then, the function $\mathbf{k}'\mathbf{b} + \boldsymbol{\lambda}'\mathbf{P}'\mathbf{b}$ is estimable in the restricted model.

Just as $\text{SSE}_r = \text{SSE}$, when the restrictions involve non-estimable functions so too, when testing the hypotheses $\mathbf{K}'\mathbf{b} = \mathbf{m}$ will $\text{SSE}_{r,H} = \text{SSE}_H$. Hence, the $F$-statistic for testing the hypothesis is identical to that of the unrestricted model. Thus, so far as calculation of the $F$-statistic is concerned, the imposition of restrictions involving non-estimable functions makes no difference at all. Both SSE and $\text{SSE}_H$ are calculated in the usual manner. Thus, the $F$-statistic is calculated just as in (71).

The fact that the model has restrictions on its parameters involving non-estimable functions does not affect the calculation of the $F$-statistic. However, these restrictions *do apply* to the hypothesis being tested, just as they do to estimable functions discussed above. Thus, hypotheses that are testable in the unrestricted model are also testable in the restricted model. However, application of the restrictions may change their form so that although they are testable in the unrestricted model. Again, consider Examples 1–17. The hypothesis $H\colon \alpha_1 - 2\alpha_2 + \alpha_3 = 17$ is testable in the unrestricted model. In a restricted model having the restriction $\alpha_1 + \alpha_3 = 3$, the hypothesis becomes $H\colon 3\alpha_1 + \alpha_3 = 23$. This hypothesis is testable in the restricted model, but is not testable in the unrestricted model.

In general, if $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is testable in the unrestricted model then, for any matrix $\mathbf{L}_{s\times q}$ $(\mathbf{K}' + \mathbf{L}\mathbf{P}')\mathbf{b} = \mathbf{m} + \mathbf{L}\boldsymbol{\delta}$ will be testable in the restricted model. It will not be testable in the unrestricted model.

The results of this section so far as estimable functions and tests of hypotheses are concerned are summarized in Tables 5.13A and 5.13B.

### c. Stochastic Constraints

In Section 6e of Chapter 3, we considered stochastic constraints of the form $\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\eta}$ where the elements of the vector $\boldsymbol{\eta}$ are independent with mean zero and variance $\boldsymbol{\tau}^2$. Again, we consider an augmented model

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix} = \begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}\mathbf{b} + \begin{bmatrix}\mathbf{e}\\ \boldsymbol{\tau}\end{bmatrix} \tag{106}$$

Where the elements of $\mathbf{e}$ are independent with mean zero and variance $\sigma^2$. However, this time, the matrices $\mathbf{X}$ and $\mathbf{R}$ and, as a result, $\begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}$ need not be of full rank. Again, we obtain the least-square estimator by minimizing

$$m = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})}{\sigma^2} + \frac{(\mathbf{r} - \mathbf{R}\mathbf{b})'(\mathbf{r} - \mathbf{R}\mathbf{b})}{\tau^2}.$$

The normal equations are

$$(\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})\hat{\mathbf{b}}^{\circ}_m = \tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}.$$

Then the mixed estimator of Theil and Goldberger (1961) takes the form

$$\hat{\mathbf{b}}^{\circ}_m = (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^{-}(\tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}), \tag{107}$$

**TABLE 5.13A Summary of Estimation in Unrestricted and Restricted Models**

| Property | Unrestricted model, and restricted model with restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$, where $\mathbf{P}'\mathbf{b}$ is non-estimable and $\mathbf{P}'$ has full-row rank $q$ | Restricted model with restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$, where $\mathbf{P}'\mathbf{b}$ is estimable and $\mathbf{P}'$ has full-row rank $q$ |
|---|---|---|
| Solutions to normal equations | $\mathbf{b}^\circ = \mathbf{G}\mathbf{X}'\mathbf{y}$ | $\mathbf{b}_r^\circ = \mathbf{b}^\circ - \mathbf{G}\mathbf{P}(\mathbf{P}'\mathbf{G}\mathbf{P})^{-1}(\mathbf{P}'\mathbf{b}^\circ - \boldsymbol{\delta})$ |
| Error sum of squares | $\mathbf{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$ | $\mathrm{SSE}_r = \mathrm{SSE} + \mathbf{t}'(\mathbf{P}'\mathbf{G}\mathbf{P})$ where $\mathbf{t}' = \mathbf{P}^1\mathbf{b}^\circ - \boldsymbol{\delta}$ |
| Estimated error variance | $\hat{\sigma}^2 = \dfrac{\mathrm{SSE}}{(N-r)}$ | $\hat{\sigma}^2 = \dfrac{\mathrm{SSE}_r}{(N-r+q)}$ |
| Estimable functions | $\mathbf{k}'\mathbf{b}$ for $\mathbf{k}'\mathbf{H} = \mathbf{k}'$ and in restricted models, $\mathbf{k}'\mathbf{b} + \boldsymbol{\lambda}'\mathbf{P}'\mathbf{b}$ for $\boldsymbol{\lambda}'$ such that $\boldsymbol{\lambda}'\boldsymbol{\delta} = 0$ (any $\boldsymbol{\lambda}'$ when $\boldsymbol{\delta} = \mathbf{0}$) | |
| Is a function that is estimable in the restricted model always estimable in the unrestricted model? | No | Yes |

**TABLE 5.13B Summary of Hypothesis Testing in Unrestricted and Restricted Models**

| Property | Unrestricted model, and restricted model with restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$, where $\mathbf{P}'\mathbf{b}$ is non-estimable and $\mathbf{P}'$ has full-row rank $q$ | Restricted model with restrictions $\mathbf{P}'\mathbf{b} = \boldsymbol{\delta}$, where $\mathbf{P}'\mathbf{b}$ is estimable and $\mathbf{P}'$ has full-row rank $q$ |
|---|---|---|
| Testable hypotheses | $\mathbf{K}'\mathbf{b} = \mathbf{m}$ for $\mathbf{K}'\mathbf{H} = \mathbf{K}'$, of full-row rank $s$ and, in restricted models, $(\mathbf{K}' + \mathbf{L}\mathbf{P}')\mathbf{b} = \mathbf{m} + \mathbf{L}\boldsymbol{\delta}$ for any $\mathbf{L}$ | |
| Is a hypothesis that is testable in the restricted model always testable in the unrestricted model? | No | Yes |
| $F$-statistic for testing testable hypotheses | $F(H) = \dfrac{(\mathbf{K}\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}\mathbf{b}^\circ - \mathbf{m})}{s\hat{\sigma}^2}$ with $(s, N-r)$ degrees of freedom | With $\mathbf{Q} = \begin{bmatrix} \mathbf{P}' \\ \mathbf{K}' \end{bmatrix}$ and $\ell = \begin{bmatrix} \boldsymbol{\delta} \\ \mathbf{m} \end{bmatrix}$ $F(H_r) = \dfrac{\text{SSE} - \text{SSE}_r + (\mathbf{Q}'\mathbf{b}^\circ - \ell)'(\mathbf{Q}'\mathbf{G}\mathbf{Q})^{-1}(\mathbf{Q}'\mathbf{b}^\circ - \ell)}{s\hat{\sigma}^2}$ with $(s, N-r+q)$ degrees of freedom. |
| Solution for $\mathbf{b}$ under null hypothesis | $\mathbf{b}^\circ_H = \mathbf{b}^\circ - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})$ | $\mathbf{b}^\circ_{r,H} = \mathbf{b}^\circ - \mathbf{G}\mathbf{Q}(\mathbf{Q}'\mathbf{G}\mathbf{Q})^{-1}(\mathbf{Q}'\mathbf{b}^\circ - \ell)$ |

where the superscripted "–" means a generalized inverse of the matrix. Using Theorem 10 of Chapter 1, we have,

$$\begin{aligned}\hat{\mathbf{b}}_m^\circ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^-(\tau^2\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{r}) \\ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^-(\tau^2\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} + \sigma^2\mathbf{R}'\mathbf{R}(\mathbf{R}'\mathbf{R})^-\mathbf{R}'\mathbf{r}) \\ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^-(\tau^2\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}_1^\circ + \sigma^2\mathbf{R}'\mathbf{R}\hat{\mathbf{b}}_2^\circ) \\ &= (\tau^2\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{R}'\mathbf{R})^-(\tau^2\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}_1 + \sigma^2\mathbf{R}'\mathbf{R}\hat{\mathbf{b}}_2)\end{aligned} \tag{108}$$

where $\mathbf{b}_1^\circ = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ and $\mathbf{b}_2^\circ = (\mathbf{R}'\mathbf{R})^-\mathbf{R}'\mathbf{r}$. In order to have unique estimators of parametric functions $\mathbf{p}'\mathbf{b}$, we need to define some different kinds of estimability.

**Definition 1** Given an augmented model in the form of (106) $\mathbf{p}'\mathbf{b}$

(i) is $\mathbf{X}$ estimable if it is estimable for the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$;
(ii) is $\mathbf{R}$ estimable if it estimable for the model $\mathbf{r} = \mathbf{R}\mathbf{b} + \tau$;
(iii) is **(X, R)** estimable if it is estimable for the model (106).

An **(X, R)** estimable function need not be $\mathbf{X}$ estimable or $\mathbf{R}$ estimable. However, an $\mathbf{X}$ or $\mathbf{R}$ estimable function is **(X, R)** estimable. This is analogous to the idea that if a hypothesis is testable in a restricted model, it may not be testable in an unrestricted model.

Observe that if a function is $\mathbf{X}$ estimable there is a $\mathbf{t}'$ where $\mathbf{p}' = \mathbf{t}'\mathbf{X} = [\mathbf{t}'\mathbf{0}]\begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}$ so that the function is **(X, R)** estimable. A similar argument applies to $\mathbf{R}$ estimable functions. Example 21 gives an example if an **(X, R)** estimable function that is not $\mathbf{X}$ estimable.

**Example 21 An (X, R) Estimable Function that is not X Estimable or R Estimable** Consider the augmented model

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix} = \begin{bmatrix}1 & 1 & 0\\ 1 & 1 & 0\\ 1 & 0 & 1\\ 1 & 0 & 1\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}b_1\\ b_2\\ b_3\end{bmatrix} + \mathbf{e}$$

where

$$\mathbf{X} = \begin{bmatrix}1 & 1 & 0\\ 1 & 1 & 0\\ 1 & 0 & 1\\ 1 & 0 & 1\end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

Assume that $\sigma^2 = \tau^2 = 1$.

Now,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}, \quad \mathbf{R}'\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{X} + \mathbf{R}'\mathbf{R} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 0 \\ 2 & 0 & 3 \end{bmatrix}.$$

The matrix $\mathbf{X}'\mathbf{X} + \mathbf{R}'\mathbf{R}$ is non-singular so every linear function is $(\mathbf{X}, \mathbf{R})$ estimable. However $b_1 + b_2 + b_3$ is neither $\mathbf{X}$ estimable or $\mathbf{R}$ estimable. A generalized inverse of

$$\mathbf{X}'\mathbf{X} \text{ is } (\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}, \quad \mathbf{H} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and

$$[1 \quad 1 \quad 1] \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = [2 \quad 1 \quad 1] \neq [1 \quad 1 \quad 1].$$

Furthermore,

$\mathbf{R}'\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is idempotent so it is its own generalized inverse and

$$[1 \quad 1 \quad 1] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = [0 \quad 1 \quad 1] \neq [1 \quad 1 \quad 1].$$

Hence $b_1 + b_2 + b_3$ is neither $\mathbf{X}$ estimable or $\mathbf{R}$ estimable. In a similar manner, we can show that $b_1 + b_2$ is $\mathbf{X}$ estimable but not $\mathbf{R}$ estimable and that $b_2$ is not $\mathbf{X}$ estimable but is $\mathbf{R}$ estimable. (See Exercise 15.) □

## 7. THE "USUAL CONSTRAINTS"

A source of difficulty with a non-full-rank model is that the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ do not have a unique solution. We have skirted this situation by using a generalized inverse of $\mathbf{X}'\mathbf{X}$. Another way to obtain a solution to the normal equations is to impose

the "usual constraints" or usual restrictions. For example, one way to solve the normal equations for the linear model

$$\mathbf{y} = \begin{bmatrix} 1_2 & 1_2 & 0 & 0 \\ 1_2 & 0 & 1_2 & 0 \\ 1_2 & 0 & 0 & 1_2 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \mathbf{e},$$

$$\begin{aligned} 6\mu^\circ + 2\alpha_1^\circ + 2\alpha_2^\circ + 2\alpha_3^\circ &= y_{..} \\ 2\mu^\circ + 2\alpha_1^\circ &= y_{1.} \\ 2\mu^\circ + 2\alpha_2^\circ &= y_{2.} \\ 2\mu^\circ + 2\alpha_3^\circ &= y_{3.} \end{aligned} \tag{109}$$

is to impose the constraint $\alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ = 0$. Using this constraint, adding the last three equations yields $6\mu^\circ = y_{..}$ and, as a result, $\mu^\circ = \bar{y}_{..}, \alpha_i^\circ = \bar{y}_{i.} - \bar{y}_{..}, i = 1, 2, 3$ as a solution. This corresponds to the solution that would be obtained using the generalized inverse

$$\mathbf{G} = \begin{bmatrix} \frac{1}{6} & 0 & 0 & 0 \\ -\frac{1}{6} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{6} & 0 & \frac{1}{2} & 0 \\ -\frac{1}{6} & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

When choosing constraints to impose, we should keep the following in mind:

1. The constraints cannot be any conditions.
2. Constraints of the form $\sum \alpha_i^\circ = 0$ are generally not the simplest for unbalanced data.
3. Constraints are not necessary to solve normal equations. They are only sufficient.
4. They can be used regardless of whether a similar relationship holds for the elements of the model.
5. In order for the solutions of the normal equations to be estimates of the parameters, there must be enough relationships to make it a full-rank model.

We will now expand on these points.

We have already seen that with any solution $\mathbf{b}^\circ$ to the normal equations, we can derive most things of interest in linear model estimation. These include $\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$, the analysis of variance, the error variance estimate $\hat{\sigma}^2 = \text{SSE}/(N - r)$, and the b.l.u.e. of any estimable function $\mathbf{k}'\mathbf{b}$ as $\hat{\mathbf{k}}'\mathbf{b} = \mathbf{k}'\mathbf{b}^\circ$. We can obtain these things

provided that we have a solution $\mathbf{b}^\circ$, no matter how it has been derived. However, for some things, we need the generalized inverse of $\mathbf{X}'\mathbf{X}$ that yielded $\mathbf{b}^\circ$. For example, the generalized inverse is, if not absolutely necessary, very helpful to ascertain the estimability of a function or to test a testable hypothesis. We shall show that applying constraints to the solutions is probably the easiest way to obtain solutions to the normal equations. However, if we want the generalized inverse corresponding to a solution to the normal equations, we must apply the constraints in a way that readily yields the generalized inverse *and* recognize the implications of doing this.

### a. Limitations on Constraints

First, the constraints need apply *only* to the elements of the solution vector $\mathbf{b}^\circ$. They are imposed solely for deriving a solution. They do not have to have anything to do with the model. Second, if the constraints are of the form $\mathbf{C}'\mathbf{b}^\circ = \gamma$, we know from (72) that minimizing $(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)$ subject to $\mathbf{C}'\mathbf{b}^\circ = \gamma$ leads to the equations

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b}^\circ + \mathbf{C}\lambda &= \mathbf{X}'\mathbf{y} \\ \mathbf{C}'\mathbf{b}^\circ &= \gamma. \end{aligned}$$

These equations are equivalent to

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \gamma \end{bmatrix}, \tag{110}$$

where $\lambda$ is a vector of Lagrange multipliers. For the equations in (110) to have one solitary solution for $\mathbf{b}^\circ$ and $\lambda$, the matrix $\mathbf{C}'$ must have full-row rank and sufficient rows to make $\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{bmatrix}$ non-singular. Applying Lemma 6 in Chapter 1 to (34), the rows of $\mathbf{C}'$ must be LIN of those of $\mathbf{X}$. That means that $\mathbf{C}'$ cannot be of the form $\mathbf{C}' = \mathbf{L}'\mathbf{X}$. Thus, the constraints $\mathbf{C}'\mathbf{b}^\circ = \gamma$ must be such that $\mathbf{C}'\mathbf{b}$ is not estimable. Therefore, they cannot be any constraints. They must be constraints for which $\mathbf{C}'\mathbf{b}$ is not estimable, and there must be $p - r$ of them where $\mathbf{X}$ has $p$ columns and rank $r$. Under these conditions, the inverse given in Section 5b of Chapter 1 can be used to obtain the unique solution of (110). This can be shown to be equivalent to the solutions obtainable by (104) and (105).

### b. Constraints of the Form $b_i^\circ = 0$

For balanced data that lead to normal equations like (109), for example, constraints of the form $\sum \alpha_i^\circ = 0$ are indeed the easiest to use. However, for unbalanced data, they are not the easiest to use. For unbalanced data, the constraints that are easiest to use are the simple ones of putting $p - r$ elements of $\mathbf{b}^\circ$ equal to zero. They cannot be any $p - r$ elements. They must be judiciously chosen to make $\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{bmatrix}$ non-singular.

We shall discuss ways of doing this in the chapters on applications (Chapters 6 and 7).

Using constraints that make some of the elements of $\mathbf{b}^\circ$ be zero is equivalent to putting those elements equal to zero in the normal equations or more exactly in $(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)'(\mathbf{y} - \mathbf{X}\mathbf{b}^\circ)$ which is minimized, subject to such constraints. This has the effect of eliminating from the normal equations, all those terms having the zeroed $b_i^{\circ\prime}$ s and also the equations corresponding to the same $b_i^{\circ\prime}$ s. This, in turn, is equivalent to eliminating from $\mathbf{X}'\mathbf{X}$, the rows and columns corresponding to those $b_i^{\circ\prime}$ s and eliminating from $\mathbf{X}'\mathbf{y}$, the corresponding elements. What remains of $\mathbf{X}'\mathbf{X}$ is a symmetric non-singular matrix of order $r$. Hence, these equations modified by the constraints of putting some $b_i^{\circ\prime}$ s zero can be solved. The solutions together with the zeroed $b_i^{\circ\prime}$ s of the constraints, then constitute $\mathbf{b}^\circ$, a solution to the normal equations. We now describe the details of this procedure and the derivation of the corresponding generalized inverse.

Putting $(p - r)$ $b_i^{\circ\prime}$ s equal to zero is equivalent to $\mathbf{C}'\mathbf{b}^\circ = \mathbf{0}$ with $\mathbf{C}'$ having $p - r$ rows each of which is null except for a single unity element. Suppose $\mathbf{R}$ is the identity matrix of order $p$ with its rows in a different sequence. Such matrices are called permutation matrices. We have that $\mathbf{R}'\mathbf{R} = \mathbf{I}$. Suppose that the permutation matrix $\mathbf{R}$ is such that

$$\mathbf{C}'\mathbf{R} = [\,\mathbf{0}_{(p-r)\times r} \quad \mathbf{I}_{p-r}\,]. \tag{111}$$

Then, remembering $\mathbf{R}$ is orthogonal, the equations to be solved (110) can be rewritten as

$$\begin{bmatrix} \mathbf{R}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{R}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b}^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{R}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \gamma \end{bmatrix}.$$

This reduces to

$$\begin{bmatrix} \mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R} & \mathbf{R}'\mathbf{C} \\ \mathbf{C}'\mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{R}'\mathbf{b}^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{R}'\mathbf{X}'\mathbf{y} \\ \mathbf{0} \end{bmatrix}. \tag{112}$$

We partition $\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R}$, $\mathbf{R}'\mathbf{b}^\circ$, and $\mathbf{R}'\mathbf{X}'\mathbf{y}$ to conform with $\mathbf{C}'\mathbf{R}$ in (111). Then,

$$\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R} = \begin{bmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} \end{bmatrix}, \quad \mathbf{R}'\mathbf{b}^\circ = \begin{bmatrix} \mathbf{b}_1^\circ \\ \mathbf{b}_2^\circ \end{bmatrix}, \quad \text{and} \quad \mathbf{R}'\mathbf{X}'\mathbf{y} = \begin{bmatrix} (\mathbf{X}'\mathbf{y})_1 \\ (\mathbf{X}'\mathbf{y})_2 \end{bmatrix}. \tag{113}$$

We then have,

$$\mathbf{Z}_{11}, \text{of full rank}, = (\mathbf{X}'\mathbf{X})_m. \tag{114}$$

We also have that

$$\mathbf{b}_1^\circ = \text{solutions of modified equations}$$

and

$$\mathbf{b}_2^\circ = \text{zeroed } b_i^{\circ\prime}\text{s}.$$

Equations (112) become

$$\begin{bmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{0} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^\circ \\ \mathbf{b}_2^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\mathbf{y})_1 \\ (\mathbf{X}'\mathbf{y})_2 \\ \mathbf{0} \end{bmatrix}.$$

Since $\mathbf{b}_2^\circ = \mathbf{0}$, the solution may be written in the form

$$\begin{bmatrix} \mathbf{b}_1^\circ \\ \mathbf{b}_2^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{11}^{-1} & \mathbf{0} & -\mathbf{Z}_{11}^{-1}\mathbf{Z}_{12} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ -\mathbf{Z}_{21}\mathbf{Z}_{11}^{-1} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} (\mathbf{X}'\mathbf{y})_1 \\ (\mathbf{X}'\mathbf{y})_2 \\ \mathbf{0} \end{bmatrix}. \tag{115}$$

The important part of this solution is

$$\mathbf{b}_1^\circ = \mathbf{Z}_{11}^{-1}(\mathbf{X}'\mathbf{y})_1. \tag{116}$$

We may derive equation (116) by multiplying the inverse of the modified $\mathbf{X}'\mathbf{X}$ matrix by the modified $\mathbf{X}'\mathbf{y}$ vector. A complete solution $\mathbf{b}^\circ$ now consists of the $\mathbf{b}_1^\circ$ and the $b_i^{\circ\prime}$ zeroed by the constraints.

We can derive the generalized inverse of $\mathbf{X}'\mathbf{X}$ corresponding to a solution (116) as follows. From (115),

$$\begin{bmatrix} \mathbf{b}_1^\circ \\ \mathbf{b}_2^\circ \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} (\mathbf{X}'\mathbf{y})_1 \\ (\mathbf{X}'\mathbf{y})_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{R}'\mathbf{X}'\mathbf{y}.$$

Using the orthogonality of $\mathbf{R}$ and (113), we obtain

$$\mathbf{b}^\circ = \mathbf{R}(\mathbf{R}'\mathbf{b}^\circ) = \mathbf{R} \begin{bmatrix} \mathbf{Z}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{R}'\mathbf{X}'\mathbf{y}. \tag{117}$$

From Section 1b of Chapter 1 with the definition of $\mathbf{Z}_{11}$ given in (114), the generalized inverse of $\mathbf{X}'\mathbf{X}$ is given by

$$\mathbf{G} = \mathbf{R} \begin{bmatrix} \mathbf{Z}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{R}' \tag{118}$$

Thus, from equation (117), $\mathbf{G}$ of (118) is the generalized inverse of $\mathbf{X}'\mathbf{X}$ corresponding to the solution $\mathbf{b}^\circ$ found by using (116) and (117). This leads to the following procedure.

### c. Procedure for Deriving b° and G

1. Find the rank of the matrix $\mathbf{X}'\mathbf{X}$ of order $p$. Call it $r$.
2. Delete $p - r$ rows and corresponding columns from $\mathbf{X}'\mathbf{X}$, to leave a symmetric sub-matrix of full rank $r$. Call the modified matrix $(\mathbf{X}'\mathbf{X})_m$.
3. Corresponding to the rows deleted from $\mathbf{X}'\mathbf{X}$ delete elements from $\mathbf{X}'\mathbf{y}$. Call the modified vector $(\mathbf{X}'\mathbf{y})_m$.
4. Calculate $\mathbf{b}_m^\circ = [(\mathbf{X}'\mathbf{X})_m]^{-1}(\mathbf{X}'\mathbf{y})_m$.
5. In $\mathbf{b}^\circ$, all elements corresponding to rows deleted from $\mathbf{X}'\mathbf{X}$ are zero. The other elements are those of $\mathbf{b}_m^\circ$ in sequence.
6. In $\mathbf{X}'\mathbf{X}$, replace all the elements of $(\mathbf{X}'\mathbf{X})_m$. by those of its inverse. Put the other elements zero. The resulting matrix is $\mathbf{G}$ its generalized inverse corresponding to the solution $\mathbf{b}^\circ$. Its derivation is in line with the algorithm of Section 1b of Chapter 1.

**Example 22 Illustration of the Procedure** Consider the linear model

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_4 & \mathbf{0} & \mathbf{1}_4 & \mathbf{0} \\ \mathbf{1}_4 & \mathbf{0} & \mathbf{0} & \mathbf{1}_4 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \mathbf{e}$$

Then,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix}$$

Step 1: The order of the matrix is $p = 4$. Its rank $r = 3$.
Steps 2 and 3: We can use any sub-matrix of rank 3 we want to. It does not have to be the one in the upper left-hand corner.

$$(\mathbf{X}'\mathbf{X})_m = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad (\mathbf{X}'\mathbf{y})_m = \begin{bmatrix} y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix}.$$

Step 4:
We find that

$$\mathbf{b}_m = \begin{bmatrix} \frac{1}{4}y_{1.} \\ \frac{1}{4}y_{2.} \\ \frac{1}{4}y_{3.} \end{bmatrix}.$$

Step 5: Putting the zero in we get

$$\mathbf{b}^\circ = \begin{bmatrix} 0 \\ \frac{1}{4}y_{1.} \\ \frac{1}{4}y_{2.} \\ \frac{1}{4}y_{3.} \end{bmatrix}.$$

Step 6: A generalized inverse is

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}$$

There are other correct solutions for this model. See how many of them you can find. □

### d. Restrictions on the Model

Throughout the preceding discussion of constraints, no mention has been made of restrictions on the parameters of the model corresponding to constraints imposed on a solution. This is because constraints on the solution are used solely for obtaining a solution and need have no bearing on the model whatever. However, if the model is such that there are restrictions on its parameters, these same restrictions can be used as constraints on the solutions, provided that they relate to non-estimable functions. More formally, this means that for restrictions $\mathbf{P}'\mathbf{b} = \delta$, $\mathbf{P}'\mathbf{b}$ is not estimable. If $\mathbf{P}'$ were of full-row rank $p - r$, then the solutions would be given by

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{P} \\ \mathbf{P}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \delta \end{bmatrix} \tag{119}$$

and the solution would in fact be the b.l.u.e. of $\mathbf{b}$. Of course, the solution to (119) could also be obtained by using the solution derived from simple constraints of the

form $b_i^\circ = 0$ discussed in subsection b, namely equation (117). This can be amended in accord with (104) and (105) to give a solution satisfying (119). The solution will be from (104)

$$\mathbf{b}_{r,0}^\circ = \mathbf{b}_0^\circ + (\mathbf{H} - \mathbf{I})\mathbf{z}_1, \tag{120}$$

using $\mathbf{b}^\circ$ of (117) as $\mathbf{b}_0^\circ$, $\mathbf{G}$ of (118). From (105), the $\mathbf{z}_1$ of (120) will be such that

$$\mathbf{P}'(\mathbf{H} - \mathbf{I})\mathbf{z}_1 = \delta - \mathbf{P}'\mathbf{G}\mathbf{X}'\mathbf{y} \tag{121}$$

as in (105). This procedure will be especially useful when the restrictions in the model $\mathbf{P}'\mathbf{b} = \delta$ involves $\mathbf{P}'$ of less than $p - r$ rows.

We have already pointed out that the important thing about restrictions in the model is their effect on estimable functions and testable hypotheses. Equally as important is the fact that constraints on the solutions do not necessarily imply restrictions on the model. Therefore, constraints do not affect estimable functions or testable hypotheses. Furthermore, since constraints are only a means of obtaining a solution $\mathbf{b}^\circ$, they do not affect sums of squares. Confusion on these points often arises because of certain kinds of restrictions that often occur. These same restrictions applied as constraints to the solution greatly aid in obtaining a solution. For example, the model equation $y_{ij} = \mu + \alpha_i + e_{ij}$ is often written as $y_{ij} = \mu_i + e_{ij}$ with $\mu$ and $\alpha_i$ defined as $\mu = \sum_{i=1}^{c} \mu_i/c$ and $\alpha_i = \mu_i - \mu$, respectively. In this way, a restriction on the model is $\sum_{i=1}^{c} \alpha_i = 0$. Suppose for $c = 3$, the normal equations were for such a model

$$\begin{aligned}
6\mu^\circ + 2\alpha_1^\circ + 2\alpha_2^\circ + 2\alpha_3^\circ &= y_{..} \\
2\mu^\circ + 2\alpha_1^\circ &= y_{1.} \\
2\mu^\circ + 2\alpha_2^\circ &= y_{2.} \\
2\mu^\circ + 2\alpha_3^\circ &= y_{3.}.
\end{aligned}$$

In order to help solve the equations and because $\alpha_1 + \alpha_2 + \alpha_3 = 0$, we impose the constraint

$$\alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ = 0. \tag{122}$$

On the other hand, suppose that the normal equations were

$$\begin{aligned}
6\mu^\circ + 3\alpha_1^\circ + 2\alpha_2^\circ + \alpha_3^\circ &= y_{..} \\
2\mu^\circ + 3\alpha_1^\circ &= y_{1.} \\
2\mu^\circ + 2\alpha_2^\circ &= y_{2.} \\
2\mu^\circ + \alpha_3^\circ &= y_{3.}.
\end{aligned} \tag{123}$$

For these normal equations, the constraint (122) is of no particular help in solving them.

A helpful constraint would be

$$3\alpha_1^\circ + 2\alpha_2^\circ + \alpha_3^\circ = 0. \tag{124}$$

However, this is no reason for making $3\alpha_1 + 2\alpha_2 + \alpha_3 = 0$ be part of the model. Not only might it be quite inappropriate, but also there is no need for it. Suppose in fact that $\alpha_1 + \alpha_2 + \alpha_3 = 0$ is a meaningful restriction in the model. Then (124) could still be used for solving equations (123). Furthermore, provided that the corresponding generalized inverse of $\mathbf{X}'\mathbf{X}$ was found, the solution could be amended to satisfy (122) by using (120) and (121). Thus, if $\mathbf{b}^\circ$ is the solution satisfying (124), then that satisfying (122) is (120) with (121) using $\mathbf{P}' = [\,0 \;\; 1 \;\; 1 \;\; 1\,]$, $\delta = 0$ and $\mathbf{G}$ corresponding to $\mathbf{b}_0^\circ$.

### e. Illustrative Examples of Results in Subsections a–d

We shall use data from Examples 1–17. Recall that from (6), the normal equations were

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \end{bmatrix} = \begin{bmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \\ y_{2\cdot} \\ y_{3\cdot} \end{bmatrix} = \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \end{bmatrix}.$$

We now give three illustrations of the procedure outlined in subsection c. In each case, we give the six steps from Subsection c. Step 1 is the same for all three illustrations, so it will not be repeated in Examples 24 and 25.

**Example 23 The First Illustration** Step 1: $p = 4$ and $r = 3$.

Steps 2 and 3: $(\mathbf{X}'\mathbf{X})_m = \begin{bmatrix} 6 & 3 & 2 \\ 3 & 3 & 0 \\ 2 & 0 & 2 \end{bmatrix}$ and $(\mathbf{X}'\mathbf{y})_m = \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \end{bmatrix}$.

Step 4: $\mathbf{b}_m^\circ = (\mathbf{X}'\mathbf{X})_m^- (\mathbf{X}'\mathbf{y})_m = \begin{bmatrix} 1 & -1 & -1 \\ -1 & \frac{4}{3} & 1 \\ -1 & 1 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \end{bmatrix} = \begin{bmatrix} 0.19 \\ 47.32 \\ 0.595 \end{bmatrix}$.

Step 5:

$$\mathbf{b}^{\circ\prime} = [\,0.19 \quad 47.32 \quad 0.595 \quad 0\,]. \tag{125}$$

Step 6: $\mathbf{G} = \begin{bmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

The value of $\mathbf{b}^\circ$ obtained above is the same as that of $\mathbf{b}_2^\circ$ in Example 2. Notice that

$$(0.19)(144.2) + (47.32)(142.53) + (1.57)(0.595) = 6772.87 = \text{SSR}.$$

The next example uses a different choice of $(\mathbf{X}'\mathbf{X})_m$. □

**Example 24 The Second Illustration** Steps 2 and 3: $(\mathbf{X}'\mathbf{X})_m = \begin{bmatrix} 6 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

and $(\mathbf{X}'\mathbf{y})_m = \begin{bmatrix} 144.29 \\ 1.57 \\ 0.19 \end{bmatrix}$.

Step 4: $\mathbf{b}_m^\circ = \frac{1}{6}\begin{bmatrix} 2 & -2 & -2 \\ -2 & 5 & 2 \\ -2 & 2 & 8 \end{bmatrix}\begin{bmatrix} 144.29 \\ 1.57 \\ 0.19 \end{bmatrix} = \begin{bmatrix} 47.51 \\ -46.725 \\ -47.32 \end{bmatrix}$.

Step 5: $\mathbf{b}^{\circ\prime} = [\,47.51 \quad -46.751 \quad -47.32\,]$.

Step 6: $\mathbf{G} = \frac{1}{6}\begin{bmatrix} 2 & 0 & -2 & -2 \\ 0 & 0 & 0 & 0 \\ -2 & 0 & 5 & 2 \\ -2 & 0 & 2 & 8 \end{bmatrix}$.

One check on this result is

$$\text{SSR} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} = (47.51)(144.29) + (-46.725)(1.57) + (-47.32)(0.19) = 6772.87$$

as before. □

The next example is the easiest computationally.

**Example 25 The Third Illustration** Step 2 and 3: $(\mathbf{X}'\mathbf{X})_m = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$,

$(\mathbf{X}'\mathbf{y})_m = \begin{bmatrix} 142.53 \\ 1.57 \\ 0.19 \end{bmatrix}$

Step 4: $\mathbf{b}_m^\circ = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 142.53 \\ 1.57 \\ 0.19 \end{bmatrix} = \begin{bmatrix} 47.51 \\ 0.785 \\ 0.19 \end{bmatrix}$

Step 5: $\mathbf{b}^{\circ\prime} = [\,0 \quad 47.51 \quad 0.785 \quad 0.19\,]$

Step 6: $\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Observe that $(47.51)(142.53) + (0.785)(1.57) + (0.19)(0.19) = 6772.87$. □

The $\mathbf{b}^\circ$ in Example 25 above is the same as $\mathbf{b}_1^\circ$ in Example 2. The sums of squares for Examples 23 and 25 are obtained in Example 6, equations (21) and (22), respectively. They are the same as that for Example 24 as the theory we are developing would predict.

**Example 26 Solution of Normal Equations with a Restriction** Suppose that the restrictions on the model are $\alpha_1 + \alpha_2 + \alpha_3 = 0$. Then, the equations (119) are

$$\begin{bmatrix} 6 & 3 & 2 & 1 & 0 \\ 3 & 3 & 0 & 0 & 1 \\ 2 & 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_2^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \\ 0 \end{bmatrix} \tag{126}$$

Inverting the $5 \times 5$ matrix the solution is

$$\begin{bmatrix} \mathbf{b}^\circ \\ \lambda \end{bmatrix} = \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \lambda \end{bmatrix} = \frac{1}{54} \begin{bmatrix} 11 & -5 & -2 & 7 & -18 \\ -5 & 17 & -4 & -13 & 18 \\ -2 & -4 & 20 & -16 & 18 \\ 7 & -13 & -16 & 29 & 18 \\ 18 & 18 & 18 & 18 & 0 \end{bmatrix} \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \\ 0 \end{bmatrix}.$$

Then,

$$\mathbf{b}^{\circ\prime} = [\, \mu^\circ \quad \alpha_1^\circ \quad \alpha_2^\circ \quad \alpha_3^\circ \,] = [\, 16.1617 \quad 31.3483 \quad -15.3767 \quad -15.9717 \,]. \tag{127}$$

An alternative way to obtain a solution is that of (120). Use a solution based on the constraint $\alpha_3^\circ = 0$ and amend it to satisfy $\alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ = 0$. To do this, use $\mathbf{b}^{\circ\prime} = [\, 0.19 \; 47.32 \; 0.595 \; 0 \,]$ of (125) where the corresponding $\mathbf{H}$ matrix is

$$\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Hence, as in (120) the solution to (126) is

$$\mathbf{b}^{\circ}_{r,0} = \mathbf{b}^{\circ\prime} = \begin{bmatrix} 0.19 \\ 47.32 \\ 0.595 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \mathbf{z}_1. \tag{128}$$

Then (121) is

$$[\,0 \quad 1 \quad 1 \quad 1\,] \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \mathbf{z}_1 = -[\,0 \quad 1 \quad 1 \quad 1\,] \begin{bmatrix} 0.19 \\ 47.32 \\ 0.595 \\ 0 \end{bmatrix}.$$

Therefore,

$$\mathbf{z}_1' = [\,z_1 \quad z_2 \quad z_3 \quad 15.971\,].$$

Substitution in (128) gives

$$\mathbf{b}^{\circ}_{r,0} = \begin{bmatrix} 0.19 \\ 47.32 \\ 0.595 \\ 0 \end{bmatrix} + 15.9717 \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 16.1617 \\ 31.3483 \\ -15.3767 \\ -15.9717 \end{bmatrix}$$

as in (127).

Suppose we use $3\alpha_1^{\circ} + 2\alpha_2^{\circ} + \alpha_3^{\circ} = 0$. The solution to

$$\begin{bmatrix} 6 & 3 & 2 & 1 & 0 \\ 3 & 3 & 0 & 0 & 1 \\ 2 & 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 3 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu^{\circ} \\ \alpha_1^{\circ} \\ \alpha_2^{\circ} \\ \alpha_2^{\circ} \\ \lambda \end{bmatrix} = \begin{bmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \\ 0 \end{bmatrix}$$

yields

$$\mathbf{b}^{\circ} = [\,24.0483 \quad 23.4617 \quad -23.2633 \quad -23.8583\,]. \tag{129}$$

The corresponding generalized inverse is

$$\mathbf{G} = \frac{1}{6}\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ -1 & 0 & 0 & 6 \end{bmatrix} \quad \text{and} \quad \mathbf{H} = \frac{1}{6}\begin{bmatrix} 1 & 0 & 2 & 1 \\ -1 & 2 & -2 & -1 \\ -1 & 0 & 4 & -1 \\ -1 & 0 & -2 & 5 \end{bmatrix}.$$

To amend this solution to satisfy $\alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ = 0$, we solve (121). For this case that is

$$[\,0 \quad 1 \quad 1 \quad 1\,]\frac{1}{6}\begin{bmatrix} 0 & 3 & 2 & 1 \\ 0 & -3 & -2 & -1 \\ 0 & -3 & -2 & -1 \\ 0 & -3 & -2 & -1 \end{bmatrix}\mathbf{z}_1 = -[\,0 \quad 1 \quad 1 \quad 1\,]\begin{bmatrix} 24.0483 \\ 23.4617 \\ -23.2633 \\ -23.8583 \end{bmatrix}$$

or

$$-(3z_2 + 2z_3 - z_4) = 47.3866.$$

Using (129) for $\mathbf{b}_0^\circ$ in (120), the solution satisfying $\alpha_1^\circ + \alpha_2^\circ + \alpha_3^\circ = 0$ is

$$\begin{aligned} \mathbf{b}_0^{r,\circ} &= \begin{bmatrix} 24.0483 \\ 23.4283 \\ -23.2633 \\ -23.8583 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} 0 & 3 & 2 & 1 \\ 0 & -3 & -2 & -1 \\ 0 & -3 & -2 & -1 \\ 0 & -3 & -2 & -1 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \\ &= \begin{bmatrix} 24.0483 \\ 23.4283 \\ -23.2633 \\ -23.8583 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} -47.3866 \\ 47.3866 \\ 47.3866 \\ 47.3866 \end{bmatrix} = \begin{bmatrix} 16.1605 \\ 31.326 \\ -15.3655 \\ -15.9605 \end{bmatrix} \end{aligned}$$

as in (127). □

## 8. GENERALIZATIONS

We have now discussed both the full-rank model and the model not of full rank. The non-full-rank model is, of course just a generalization of the full-rank model. As has already been pointed out the full-rank model is a special case of the non-full-rank model with $\mathbf{G}$ and $\mathbf{b}^\circ$ taking the forms $(\mathbf{X}'\mathbf{X})^{-1}$and $\hat{\mathbf{b}}$, respectively. Therefore, in general, the non-full-rank model covers all the cases.

Estimability and testability, however, only enter into the non-full-rank model. For the full-rank case, all linear functions are testable and all linear hypotheses are testable. Therefore, there is merit in dealing with the two models separately, as we have done.

However, in both models, only one special case has been considered concerning the variance of the error terms in the linear model. This is the case where the error terms have $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. We now briefly consider the general case of $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$, both where $\mathbf{V}$ is non-singular and where $\mathbf{V}$ is singular.

### a. Non-singular V

When $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ with $\mathbf{V}$ non-singular, the normal equations are as indicated in Section 3 of Chapter 3.

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{130}$$

For the full-rank model, the normal equations in (130) have the single solution

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{131}$$

as given in Section 3 of Chapter 3. For the non-full-rank model, a generalized inverse of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ must be used to solve (130). If we denote this by $\mathbf{F}$, we obtain

$$\hat{\mathbf{b}} = \mathbf{F}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad \text{with} \quad \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{F}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}. \tag{132}$$

The result in (131) is a special case of that in (132). We thus see that estimation in the model using $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ for non-singular $\mathbf{V}$ is identical to that when $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$ with the following exceptions. First, we use a generalized inverse of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ instead of an ordinary inverse. Second, we use $\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ in place of $\mathbf{X}'\mathbf{y}$.

Furthermore, since $\mathbf{V}$ is a symmetric positive definite matrix, there exists a non-singular $\mathbf{L}$ such that $\mathbf{V}^{-1} = \mathbf{L}\mathbf{L}'$. Putting $\mathbf{x} = \mathbf{L}'\mathbf{y}$ transforms the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ into $\mathbf{x} = \mathbf{L}'\mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = \mathbf{L}'\mathbf{e}$ and $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Estimating $\mathbf{b}$ from this model for $\mathbf{x}$ gives $\hat{\mathbf{b}}$ or $\mathbf{b}^\circ$ from (131) or (132), respectively. The corresponding error sum of squares is

$$\mathbf{x}'\mathbf{x} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{L}\mathbf{x} = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{b}^\circ\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{133}$$

In the full-rank case, we use $\hat{\mathbf{b}}$ for $\mathbf{b}^\circ$. Thus, we use the weighted sum of squares $\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$ in place of $\mathbf{y}'\mathbf{y}$ in the corresponding analysis of variance.

### b. Singular V

At least two conditions among data can lead to $\text{var}(\mathbf{y}) = \mathbf{V}$ being singular. One condition is when any elements of $\mathbf{y}$ are linear functions of other elements. Another

is if any elements of $\mathbf{y}$ are a constant plus linear functions of other elements. For example, if,

$$v(y_1) = v(y_2) = \sigma^2 \quad \text{and} \quad \text{cov}(y_1, y_2) = 0$$

then,

$$\text{var}\begin{bmatrix} y_1 \\ y_2 \\ y_1 + y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}\sigma^2; \tag{134}$$

and for any constant $\theta$,

$$\text{var}\begin{bmatrix} y_1 \\ y_2 \\ y_1 + \theta \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}\sigma^2. \tag{135}$$

Suppose we write $\mathbf{w}'$ for the vector $[\,y_1 \quad y_2\,]$, and let the equation of the model be

$$\mathbf{w} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{Tb} + \boldsymbol{\varepsilon}. \tag{136}$$

Then the equation for

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_1 + y_2 \end{bmatrix}$$

of (134) can be written as

$$\mathbf{y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}\mathbf{w} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}(\mathbf{Tb} + \boldsymbol{\varepsilon}),$$

that is, as $\mathbf{y} = \mathbf{Mw}$ for $\mathbf{w}$ of (136). On the other hand, the equation for

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_1 + \theta \end{bmatrix}$$

of (135) cannot be written as $\mathbf{y} = \mathbf{M}\mathbf{w}$ for $\mathbf{w}$ of (136). We only consider the case where $\mathbf{y}$ can be written in the form $\mathbf{y} = \mathbf{M}\mathbf{w}$ like (134). Zyskind and Martin (1969) consider the more general case where $\mathbf{V}$ is singular but $\mathbf{y}$ cannot necessarily be written in the form $\mathbf{y} = \mathbf{M}\mathbf{w}$. The situation where $\mathbf{y} = \mathbf{M}\mathbf{w}$ is a special case of their results. We consider it, briefly, because it is the way a singular $\mathbf{V}$ frequently arises. The normal equations, their solutions, and the ensuing results are most easily described for this case.

Whenever some elements of $\mathbf{y}$ can be described as functions of other elements, $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{M}\mathbf{w}, \tag{137}$$

where no element of $\mathbf{w}$ is a linear function of the others. Thus, $\mathbf{M}$ has full-column rank. Furthermore, on taking the equation of the model for $\mathbf{w}$ as being

$$\mathbf{w} = \mathbf{T}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{138}$$

we have

$$\mathbf{y} = \mathbf{M}\mathbf{w} = \mathbf{M}\mathbf{T}\mathbf{b} + \mathbf{M}\boldsymbol{\varepsilon}.$$

As a result, if the model for $\mathbf{y}$ is $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, we can take

$$\mathbf{X} = \mathbf{M}\mathbf{T} \tag{139}$$

and $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$. Furthermore, if $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and $\text{var}(\mathbf{y}) = \mathbf{V}\sigma^2$, we have

$$\mathbf{V}\sigma^2 = \text{var}(\mathbf{y}) = \text{var}(\mathbf{e}) = \text{var}(\mathbf{M}\boldsymbol{\varepsilon}) = \mathbf{M}\mathbf{M}'\sigma^2$$

so that

$$\mathbf{V} = \mathbf{M}\mathbf{M}'. \tag{140}$$

From (136), the normal equations for $\mathbf{b}^\circ$ are

$$\mathbf{T}'\mathbf{T}\mathbf{b}^\circ = \mathbf{T}'\mathbf{w}. \tag{141}$$

However, since $\mathbf{M}$ has full-column rank it can be readily shown that $\mathbf{M}'(\mathbf{M}'\mathbf{M})^{-2}\mathbf{M}'$ is the unique Moore–Penrose generalized inverse of $\mathbf{V}$ of (140) by checking that it satisfies all four of the defining axioms. Furthermore, by Theorem 10 of Chapter 1, $\mathbf{M}'(\mathbf{M}\mathbf{M}')^{-}\mathbf{M}$ is unique for all generalized inverses $\mathbf{V}^{-} = (\mathbf{M}\mathbf{M}')^{-}$of $\mathbf{V} = \mathbf{M}\mathbf{M}'$. Using the Moore–Penrose inverse for this shows that

$$\mathbf{M}'\mathbf{V}^{-}\mathbf{M} = \mathbf{M}'(\mathbf{M}\mathbf{M}')^{-}\mathbf{M} = \mathbf{M}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-2}\mathbf{M}'\mathbf{M} = \mathbf{I}. \tag{142}$$

Rewrite (141) as $\mathbf{T}'\mathbf{I}\mathbf{T}\mathbf{b}^\circ = \mathbf{T}'\mathbf{I}\mathbf{w}$. Using (142), this becomes $\mathbf{T}'\mathbf{M}'\mathbf{V}^-\mathbf{M}\mathbf{T}\mathbf{b}^\circ = \mathbf{T}'\mathbf{M}'\mathbf{V}^-\mathbf{M}\mathbf{w}$.

From (137) and (139), this is equivalent to

$$\mathbf{X}'\mathbf{V}^-\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{V}^-\mathbf{y}. \tag{143a}$$

Hence,

$$\mathbf{b}^\circ = (\mathbf{X}'\mathbf{V}^-\mathbf{X})^-\mathbf{X}'\mathbf{V}^-\mathbf{y}, \tag{143b}$$

where $(\mathbf{X}'\mathbf{V}^-\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}'\mathbf{V}^-\mathbf{X}$ and $\mathbf{V}^-$ is any generalized inverse of $\mathbf{V}$. The results obtained in (143) are identical to those for non-singular $\mathbf{V}$, (130) and (131), only with a generalized inverse $\mathbf{V}^-$ of $\mathbf{V}$ used in place of $\mathbf{V}^{-1}$.

For the singular case from fitting (138) is

$$\text{SSE} = \mathbf{w}'\mathbf{w} - \mathbf{b}^{\circ\prime}\mathbf{T}'\mathbf{w}. \tag{144}$$

With the aid of (137), (139), and (142), in the same way that (143) was derived, (144) reduces to

$$\text{SSE} = \mathbf{y}'\mathbf{V}^-\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{V}^-\mathbf{y}.$$

This is the same result as (133) using $\mathbf{V}^-$ in place of $\mathbf{V}^{-1}$. From (144), its expected value is

$$\begin{aligned} E(\text{SSE}) &= E(\mathbf{w}'\mathbf{w} - \mathbf{b}^{\circ\prime}\mathbf{T}'\mathbf{w}) \\ &= [(\text{number of elements in } \mathbf{w}) - r(\mathbf{T})]\sigma^2 \\ &= [r(\mathbf{M}) - r(\mathbf{T})]\sigma^2 \end{aligned}$$

Since $\mathbf{M}$ has full-column rank using (140) and (139), we see that this is equivalent to $E(\text{SSE}) = [r(\mathbf{V}) - r(\mathbf{X})]\sigma^2$.

Hence an unbiased estimator of $\sigma^2$ is

$$\sigma^2 = \frac{\text{SSE}}{r(\mathbf{V}) - r(\mathbf{X})} = \frac{\mathbf{y}'\mathbf{V}^-\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{V}^-\mathbf{y}}{r(\mathbf{V}) - r(\mathbf{X})}.$$

Another somewhat different treatment of this topic is available in C.R. Rao (1973).

## 9. AN EXAMPLE

Throughout this chapter, we have illustrated the ideas that we presented using six data points for DNA content of three different crustaceans. We now present 31 data points and do some analyses using SAS and at the same time illustrate a few of the ideas presented in this chapter.

**Example 27 DNA content of Three Different Crustaceans** The data are presented in Table 5.14.

| Amphipods | Barnacles | Branchiopods |
|---|---|---|
| 0.74 | 0.67 | 0.19 |
| 0.95 | 0.90 | 0.21 |
| 1.71 | 1.23 | 0.22 |
| 1.89 | 1.40 | 0.22 |
| 3.80 | 1.56 | 0.28 |
| 3.97 | 2.60 | 0.30 |
| 7.16 | | 0.40 |
| 8.48 | | 0.47 |
| 13.49 | | 0.63 |
| 16.09 | | 0.87 |
| 27.00 | | 2.77 |
| 50.91 | | 2.91 |
| 64.62 | | |

The linear model is

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_{13} & \mathbf{1}_{13} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{6} & \mathbf{0} & \mathbf{1}_{6} & \mathbf{0} \\ \mathbf{1}_{12} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{12} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \mathbf{e}$$

We have that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 31 & 13 & 6 & 12 \\ 13 & 13 & 0 & 0 \\ 6 & 0 & 6 & 0 \\ 12 & 0 & 0 & 12 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{13} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{12} \end{bmatrix} \text{ and } \mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

The function $\mathbf{q}'\mathbf{b}$ is estimable if for $\mathbf{q}' = [\, q_1 \;\; q_2 \;\; q_3 \;\; q_2 \,]$, $\mathbf{q}'\mathbf{H} = \mathbf{q}$ or $q_1 = q_2 + q_3 + q_4$.

Consider the estimable function $\alpha_1 - \alpha_3$. To find an estimable function that is orthogonal to it, we solve the equation

$$[\, 0 \quad 1 \quad 0 \quad -1 \,] \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{13} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{12} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = 0.$$

One case where this is true and the function is estimable is when a = 0, b = 12, c = −25, and d = 13. The orthogonal estimable parametric function is $12\alpha_1 - 25\alpha_2 + 13\alpha_3$.

In the SAS output below, we do the analysis of variance to determine whether there is a significant difference in the average DNA content of the three types of crustaceans. We shall also test the hypotheses

$$H_{01}: \alpha_1 - \alpha_2 = 0$$
$$H_{02}: 13\alpha_1 - 25\alpha_2 + 12\alpha_3 = 0$$
$$H_{03}: \alpha_2 - \alpha_3 = 0$$
$$H_{04}: \alpha_1 - \alpha_3 = 0$$

and interpret the results where appropriate.

| **Class Level Information** | **F** | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **crust** | 3 | 1 2 3 |

| | | | | | |
|---|---|---|---|---|---|
| **Number of Observations Read** | 31 | | | | |
| **Number of Observations Used** | 31 | | | | |

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 2 | 1580.104795 | 790.052397 | 4.42 | 0.0215 |
| **Error** | 28 | 5009.667102 | 178.916682 | | |
| **Corrected Total** | 30 | 6589.771897 | | | |

| **R-Square** | **Coeff Var** | **Root MSE** | **amt Mean** |
|---|---|---|---|
| 0.239781 | 189.7388 | 13.37597 | 7.049677 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **crust** | 2 | 1580.104795 | 790.052397 | 4.42 | 0.0215 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **crust** | 2 | 1580.104795 | 790.052397 | 4.42 | 0.0215 |

| **Contrast** | **DF** | **Contrast SS** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **1 vs 3** | 1 | 1340.662895 | 1340.662895 | 7.49 | 0.0106 |
| **orthogonal to 1 vs 3** | 1 | 239.441899 | 239.441899 | 1.34 | 0.2571 |
| **1 vs 2** | 1 | 812.727632 | 812.727632 | 4.54 | 0.0420 |
| **2 vs 3** | 1 | 1.380625 | 1.380625 | 0.01 | 0.9306 |

From the SAS output we see that there is a significant difference amongst the DNA content of the three crustaceans. The DNA content of amphiboids is significantly different from both barnacles and branchiopods. However, there is not a statistically significant difference between barnacles and branchiopods. The contrast orthogonal to $\alpha_1 - \alpha_3$ is not significantly different from zero.

The code used to generate the above output follows:

```
data dna;
input crust amt;
cards;
1 .74
.............
3 2.91
proc glm;
class crust;
model amt=crust;
contrast '1 vs 3' crust 1 0 -1;
contrast 'orthogonal to 1 vs 3' crust 13 -25 12;
contrast '1 vs 2' crust 1 -1 0;
contrast '2 vs 3' crust 0 1 -1;
run;
```

□

## 10. SUMMARY

The basic results of this chapter are summarized at the beginning of the next, before using them on applications in that and succeeding chapters. Additional summaries are to be found as follows:

| | | |
|---|---|---|
| Procedure for deriving **G** | : | Section 7c. |
| Analysis of variance for fitting model | : | Tables 5.5 and 5.6, Section 3g. |
| Estimable functions | : | Table 5.8, Section 4d. |
| Analysis of variance for testing hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{0}$ | : | Tables 5.9 and 5.10, Section 5c. |
| Restricted models | : | Tables 5.13A and 5.13B, Section 6. |

## 11. EXERCISES

**1** The matrices

$$\mathbf{G}_3 = \frac{1}{96}\begin{bmatrix} 11 & -3 & 1 & 13 \\ -3 & 27 & -9 & -21 \\ 1 & -9 & 35 & -25 \\ 13 & -21 & -25 & 59 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_4 = \frac{1}{6}\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ -1 & 0 & 0 & 6 \end{bmatrix}$$

are generalized inverses of $\mathbf{X}'\mathbf{X}$ in (3). For the data of Table 5.1

**(a)** Find the solutions to the normal equations.

**(b)** Find two linear combinations of these solutions that are the same and two that are different.

**(c)** Show that $\hat{\mathbf{y}}$ and SSR are the same. For both generalized inverses show that $\hat{\mathbf{y}}$ and SSR are the same as those obtained in Example 3.

**(d)** Show that $\alpha_2^\circ - \alpha_3^\circ$ has the same variance when derived from $\mathbf{G}_3$ and $\mathbf{G}_4$.

**(e)** For the data of Example 1, obtain $\hat{\sigma}^2$ by using $\mathbf{G}_3$ and $\mathbf{G}_4$.

**2** For the examples that pertain to data in Table 5.1, derive the contrasts specified below and find the numerator sum of squares for testing the hypotheses that these contrasts are zero. Define orthogonal as in (93).

**(a)** A contrast orthogonal to both $6\mu + 3\alpha_1 + 2\alpha_2 + \alpha_3$ and $\alpha_1 - 2\alpha_2 + \alpha_3$.

**(b)** Two contrasts orthogonal to one another and $\alpha_1 - \alpha_2$.

**(c)** For each of the contrasts in (a) and (b), find the sum of squares due to each hypothesis that they are zero and the reduced sum of squares of Table 5.9. Show that the sum of squares due to the orthogonal contrasts add up to the regression sum of squares.

**3** For the data for Example 27, find a contrast orthogonal to each of $\alpha_2 - \alpha_3 = 0$ and $\alpha_1 - \alpha_2 = 0$. Find the sums of squares associated with these contrasts and test for statistical significance. Show that the sum of squares associated with the given contrast and the one orthogonal to it add up to the regression sum of squares.

**4** The life lengths of four different brands of light bulbs are being compared. The results follow.

| A | B | C | D |
|---|---|---|---|
| 915 | 1011 | 989 | 1055 |
| 912 | 1001 | 979 | 1048 |
| 903 | 1003 | 1004 | 1061 |
| 893 | 992 | 992 | 1068 |
| 910 | 981 | 1008 | 1053 |
| 890 | 1001 | 1009 | 1063 |
| 879 | 989 | 996 | |
| | 1003 | 998 | |
| | | 997 | |

**(a)** Set up the linear model and find the normal equations.

**(b)** Solve the normal equations by using the constraint $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$. What generalized inverse corresponds to the use of this constraint?

**(c)** Formulate the ANOVA table.

**(d)** By formulating appropriate contrasts and testing hypotheses about them, determine

whether there is a statistically significant difference between the average life length for

**(1)** A and B

**(2)** B and C

**(3)** C and D

**(4)** The average of A, B and the average of C and D.

**5** If $\mathbf{T}$ has full-row rank, prove that $\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}' = \mathbf{I}$.

**6** Show, formally, that testing the hypothesis $\lambda\mathbf{K}'\mathbf{b} = \mathbf{0}$ is identical to testing $\mathbf{K}'\mathbf{b} = \mathbf{0}$ for $\lambda$ being a scalar.

**7** Show using the notation for the singular value decompositions of $\mathbf{X}$ from Chapter 1 that $\text{SSR} = \mathbf{y}'\mathbf{S}'\mathbf{S}\mathbf{y}$ and $\text{SSE} = \mathbf{y}'\mathbf{T}'\mathbf{T}\mathbf{y}$.

**8** Show that for estimable functions $\mathbf{p}'\mathbf{b}$, where $H$ istestable, $\mathbf{p}'\mathbf{b}^{\circ}_{H}$ is independent of the generalized inverse $\mathbf{G}$.

**9** Consider the reparametization of the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ to the full-rank model $\mathbf{y} = \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{b} + \mathbf{e} = \mathbf{X}\mathbf{U}\mathbf{g} + \mathbf{e}$, where $\mathbf{g} = \mathbf{U}'\mathbf{b}$, Show that

**(a)** The least-square estimator for the reparametized model is $\hat{\mathbf{g}} = \Lambda^{-1}\mathbf{U}'\mathbf{X}'\mathbf{y}$.

**(b)** When $H$: $\mathbf{K}'\mathbf{b} = \mathbf{m}$ is testable, there exists a $\mathbf{C}'$ such that $\mathbf{K}' = \mathbf{C}'\mathbf{U}'$. Show that the hypothesis reduces to $\mathbf{C}'\mathbf{g} = \mathbf{m}$.

**(c)** In terms of the reparametized model, show that the equivalent of equation (117) of Chapter 3 is $\hat{\mathbf{g}}_H = \hat{\mathbf{g}} - \Lambda^{-1}\mathbf{C}(\mathbf{C}'\Lambda^{-1}\mathbf{C})^{-1}(\mathbf{C}'\hat{\mathbf{g}} - \mathbf{m})$.

**(d)** Using $\mathbf{K}' = \mathbf{C}'\mathbf{U}'$ and the fact that for any generalized inverse G of $\mathbf{X}'\mathbf{X}$, $\mathbf{U}'\mathbf{G}\mathbf{U} = \Lambda^{-1}$, show that $\hat{\mathbf{g}}_H = \mathbf{U}'\mathbf{b}^{\circ}_{H}$ where $\mathbf{b}^{\circ}_{H}$ is that obtained in (74). Then for estimable functions, we have that $\mathbf{p}'\mathbf{b}^{\circ}_{H} = \mathbf{p}'\mathbf{b}^{\circ} - \mathbf{p}'\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^{\circ} - \mathbf{m})$ independent of the choice of $\mathbf{G}$.

**10** Let $\mathbf{K}'\mathbf{b} = \mathbf{m}$ be a testable hypothesis. Reformulate the optimization problem in (72) as that of finding $\mathbf{b}^{\circ}_{H}$ and $\theta$ as the solution to the matrix equation.

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_H \\ \theta \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{m} \end{bmatrix}.$$

Show how the results in equations (73) and (74) may be obtained using the formula for the generalized inverse of a partitioned matrix.

**11** For $\mathbf{X}$ of order $N \times p$ and rank $r$, and $\mathbf{S}'$ and $\mathbf{S}'\mathbf{X}'$ of full-row rank $r$, show that $\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'$ is a reflexive generalized inverse of $\mathbf{X}'\mathbf{X}$.

**12** Suppose a model can be expressed as

$$y_{ijk} = \alpha_i + \varepsilon_{ijk},$$

where $y_{ijk}$ is an observation and $i = 1, \ldots, c, j = 1, \ldots, N_i,$ and $k = 1, \ldots, n_{ij}$. The vector of observations can be written as

$$\mathbf{y}' = [y_{111}\ y_{112} \cdots y_{11n_{11}} \cdots y_{1N_1 1} \cdots y_{1,N_1,n_{1N_1}} \cdots y_{c,N_c,1} \cdots y_{c,N_c,n_{cN_c}}],$$

where the observations are ordered by $k$, within $j$ within $i$. If $\mathbf{V}$ is the variance–covariance matrix of $\mathbf{y}$, it is a diagonal matrix of matrices $\mathbf{A}_{ij}$ , for $i = 1, \ldots, c$ and $j = 1, \ldots, N_i$, where $\mathbf{A}_{ij} = e\mathbf{I}_{n_{ij}} + b\mathbf{J}_{n_{ij}}$, and $\mathbf{1}'_{n_{ij}}$ is a vector of $n_{ij}$ $\mathbf{1}$'s and $\mathbf{J}_{n_{ij}} = \mathbf{1}_{n_{ij}}\mathbf{1}'_{n_{ij}}$. The normal equations for estimating $\alpha$, the vector of the $\alpha_i$'s are then $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\alpha = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, where $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ exists.

**(a)** For $c = 2$ with $n_{11} = 2, n_{12} = 3, n_{21} = 4, n_{22} = 1,$ and $n_{23} = 2$, write down $\mathbf{y}'$ and $\mathbf{V}$ in full.

**(b)** For the general case write down $\mathbf{X}$ and $\mathbf{V}$.

**(c)** Solve the normal equations for $\hat{\alpha}$, showing that

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{N_i} \dfrac{\bar{y}_{ij}}{b + e/n_{ij}}}{\sum_{j=1}^{N_i} \dfrac{1}{b + e/n_{ij}}}.$$

**13** Using

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{0} \end{bmatrix}$$

given in Section 5b of Chapter 1, show that the resulting solutions of equations (77) are $\theta = \mathbf{0}$ and $\mathbf{b}^{\circ}_{H}$ of (80) as obtained in this chapter. (We represent the matrix $\mathbf{H}$ of Section 5b of Chapter 1 by $\mathbf{K}'$ of the non-testable hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ with $\mathbf{K}'$ of full-row rank $p - r$; and $m$ of Section 5b of Chapter 1 is $p - r$ here.)

**14** Verify the result of equation (100).

**15** In Example 21, show that $b_1 + b_2$ is $\mathbf{X}$ estimable but not $\mathbf{R}$ estimable and that $b_2$ is not $\mathbf{X}$ estimable but is $\mathbf{R}$ estimable.

# 6

# TWO ELEMENTARY MODELS

We now demonstrate the methods of the preceding chapter for specific applications. We shall consider unbalanced data in detail with passing reference to the simpler cases of balanced data. The applications we shall discuss do by no means exhaust the great variety available. However, they cover a sufficiently wide spectrum for the reader to gain an adequate understanding of the methodology. He/she may apply what he/she has learned to other situations.

Throughout this and the next two chapters, we shall assume that the individual error terms have mean zero and variance $\sigma^2$ and are pairwise uncorrelated. In symbols, we assume that $E(\mathbf{e}) = \mathbf{0}$ and $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. For purposes of point estimation, these are the only assumptions that we need. However, for hypothesis testing and confidence interval estimation, we assume in addition that the error terms are normally distributed. Thus, for point estimation, we assume that $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. For hypothesis testing and confidence intervals, we assume that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. A more general assumption would be $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ for $\mathbf{V}$ symmetric and positive definite (or perhaps positive semi-definite). Although there is a brief discussion of this in Section 8 of Chapter 5, we will postpone examples of the use of the more general assumption to Chapters 9 and 10 under the heading of "mixed models."

Some of the numerical illustrations will be based on hypothetical data with numbers chosen to simplify the arithmetic. This is particularly useful for illustrating the use of formulae that arise in presenting the methodology. We shall also use some real data illustrating the results with computer outputs using either R or SAS.

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

## 1. SUMMARY OF THE GENERAL RESULTS

For ready reference, we summarize the main results of Chapter 5 that are used in this and the next two chapters.

The equation of the model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}. \tag{1}$$

The normal equations for this model are

$$\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y} \tag{2}$$

The solution to the normal equations takes the form

$$\mathbf{b}^\circ = \mathbf{G}\mathbf{X}'\mathbf{y}, \tag{3}$$

where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Recall that this means that $\mathbf{G}$ satisfies

$$\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}.$$

Development of the general theory in Chapter 5 has, as its starting point, the finding of the matrix $\mathbf{G}$. However, Section 7b of Chapter 5 describes a procedure for solving the normal equations by putting some elements of $\mathbf{b}^\circ$ equal to zero and then finding the $\mathbf{G}$ that corresponds to this solution. In certain cases, this is an easy procedure. Putting some elements of $\mathbf{b}^\circ$ equal to zero so greatly simplifies the normal equations that their solution becomes "obvious," and the corresponding $\mathbf{G}$ (by the methods of Section 5c) equally so. The basis of this procedure when $\mathbf{X}'\mathbf{X}$ has order $p$ and rank $r$ is to

$$\text{set } p - r \text{ elements of } \mathbf{b}^\circ \text{equal to zero} \tag{4}$$

and to strike out the corresponding equations of the normal equations, leaving a set of $r$ equations of full rank. Details are given in Section 7 of Chapter 5.

Once we obtain a value for $\mathbf{b}^\circ$, we see that the predicted value of $\mathbf{y}$ corresponding to its observed value is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} \tag{5}$$

The residual sum of squares is

$$\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}.$$

The estimated error variance is

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{N - r}, \text{ where } r = r(\mathbf{X}). \tag{6}$$

The sum of squares due to fitting the mean is

$$\text{SSM} = N\bar{y}^2, \tag{7}$$

where $\bar{y}$ is the mean of all observations. The sum of squares due to fitting the model is

$$\text{SSR} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}. \tag{8}$$

The total sum of squares is

$$\text{SST} = \mathbf{y}'\mathbf{y} = \sum y^2 \tag{9}$$

where $\sum y^2$ represents the sum of squares of the individual observations. Hence

$$\text{SSE} = \text{SST} - \text{SSR}. \tag{10}$$

Furthermore, SSR and SST both corrected for the mean are

$$\text{SSR}_\text{m} = \text{SSR} - \text{SSM} \tag{11}$$

and

$$\text{SST}_\text{m} = \text{SST} - \text{SSM} \tag{12}$$

with

$$\text{MSR}_\text{m} = \frac{\text{SSR}_\text{m}}{(r-1)}.$$

These calculations are summarized in the "Analysis of Variance" Tables 5.5 and 5.6 of Section 3 of Chapter 5. From them comes the coefficient of determination

$$R^2 = \frac{\text{SSR}_\text{m}}{\text{SST}_\text{m}} \tag{13}$$

In addition, on the basis of normality,

$$F(R_\text{m}) = \frac{\text{MSR}_\text{m}}{\text{MSE}} \tag{14}$$

compared to the tabulated values of the $F_{r-1,N-r}$-distribution tests whether the model $E(y) = \mathbf{Xb}$ over and above the general mean, accounts for variation in the $y$ variable. Similarly,

$$F(M) = \frac{\text{SSM}}{\text{SSE}} = \frac{N\bar{y}^2}{\hat{\sigma}^2} \tag{15}$$

compared to tabulated values of $F_{1,N-r}$ tests the hypothesis $H$: $E(\bar{y}) = 0$. An identical test is the comparison of $\sqrt{F(M)}$ against the $t_{N-r}$-distribution.

In Section 4 of Chapter 5, we saw that

1. The expected value of any observation is estimable. This means that every element of **Xb** is estimable.
2. The b.l.u.e. (best linear unbiased estimator) of any element of **Xb** (best linear unbiased estimator) is the same element of $\mathbf{Xb}^\circ$.
3. Any linear combination of elements of Xb is estimable. Its b.l.u.e. is the same linear combination of elements of **Xb.**

More generally,

$$\mathbf{q'b} \text{ is estimable when } \mathbf{q'} = \mathbf{t'X} \text{ for any } \mathbf{t'}. \tag{16}$$

As a result,

$$\widehat{\mathbf{q'b}} = \mathbf{q'b}^\circ \text{ is the b.l.u.e. of } \mathbf{q'b} \tag{17}$$

with

$$\text{var}(\widehat{\mathbf{q'b}}) = \mathbf{q'Gq}\sigma^2. \tag{18}$$

The $100(1-\alpha)$ % symmetric confidence interval on $\mathbf{q'b}$ is

$$q'\mathbf{b}^\circ \pm \hat{\sigma} t_{N-r,\frac{1}{2}\alpha} \sqrt{\mathbf{q'Gq}}. \tag{19}$$

Table 5.7 in Chapter 5 shows a variety of special cases of estimable functions. A test of the general linear hypothesis

$$H\text{: } \mathbf{K'b} = \mathbf{m}, \text{ for } \mathbf{K'b} \text{ estimable and } \mathbf{K'} \text{ having full-row rank } s \tag{20}$$

is to compare

$$F(H) = \frac{Q}{s\sigma^2}, \tag{21}$$

where,

$$Q = (\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}),$$

against tabulated values of the $F_{s,N-r}$-distribution. The solution to the normal equations under the null hypothesis is then, if needed,

$$\mathbf{b}_H^\circ = \mathbf{b}^\circ - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m}).$$

Of particular interest are hypotheses of the form $\mathbf{K}'\mathbf{b} = \mathbf{0}$ where $\mathbf{m}$ of the general case in (20) is null. Such hypotheses are discussed in Section 5c of Chapter 5. This section also contains the analysis of variance table and the appropriate $F$-tests. Section 5g of Chapter 5 deals with orthogonal contrasts $\mathbf{k}_i'\mathbf{b}$ among the elements of $\mathbf{b}$. These contrasts have the property

$$\mathbf{k}_i'\mathbf{G}\mathbf{k}_j = 0 \quad \text{for} \quad i \neq j. \tag{22}$$

When (22) is true for $i, j = 1, 2, \ldots, r$, the test of the hypothesis $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$ has a numerator sum of squares that not only equals SSR. It also equals the sum of the numerator sums of squares for testing the $r$ hypotheses $H_i$: $\mathbf{k}_i'\mathbf{b} = \mathbf{0}$, where $\mathbf{K}' = \{\mathbf{k}_i'\}$ for $i = 1, 2, \ldots, r$.

Chapter 5 also deals with models that include restrictions on the parameters. Their analyses are summarized in Table 5.13 of Chapter 5.

## 2. THE ONE-WAY CLASSIFICATION

Chapter 4 contains discussion of data about the investment in consumer durables of people with different levels of education. Assume that investment is measured by an index number. Suppose that available data consist of seven people as shown in Table 6.1. This is a very small example. However, it is adequate for purposes of illustration.

### a. The Model

Section 3 of Chapter 4 suggests the following suitable model for these data,

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{23}$$

**TABLE 6.1 Investment Indices of Seven People**

| Level of Education | No. of People | Indices | Total |
|---|---|---|---|
| 1. High school incomplete | 3 | 74, 68, 77 | 219 |
| 2. High school graduate | 2 | 76, 80 | 156 |
| 3. College graduate | 2 | 85, 93 | 178 |
| | 7 | | 553 |

The dependent variable $y_{ij}$ is the investment index of the $j$th person in the $i$th education level. The term $\mu$ is a general mean. The effect of the $i$th level of education is represented by $\alpha_i$. The $e_{ij}$ represents the random error term peculiar to $y_{ij}$. For the data of Table 6.1, there are three education levels. Thus, $i$ takes values $i = 1, 2, 3$. For a given $i$, the subscript $j$ takes values $j = 1, 2, \ldots, n_i$, where $n_i$ is the number of observations in the $i$th education level. For this example, from Table 6.1, we have $n_1 = 3, n_2 = 2,$ and $n_3 = 2$.

The model (23) is the model for the one-way classification. In general, the groupings such as education levels are called classes. In (23), $y_{ij}$ is the effect of the response of the $i$th class, $\mu$ is a general mean, $\alpha_i$ is the effect of the response of the $i$th class and $e_{ij}$ is the error term. When the number of classes in the data is $a$, $i = 1, 2, \ldots, a$, with $j = 1, 2, \ldots, n_i$. Although described here in terms of investment as the response and level of education as the classes, this type of model can apply to many situations. For example, the classes may be varieties of a plant, makes of a machine, or different levels of income in the community. The word "treatment" is sometimes used instead of "classes." For example, if we wish to compare the effects of different fertilizers on the yield of corn, say, we might consider the fertilizer used as a treatment and use the same kind of model. Analysis of this model has already been used in Chapter 5, interspersed with the development of the general methods of that chapter. We give a further example here and indicate some results that apply to the model generally.

The normal equations come from writing the data of Table 6.1 in terms of equation (23). We have that

$$\begin{bmatrix} 74 \\ 68 \\ 77 \\ 76 \\ 80 \\ 85 \\ 93 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 + e_{11} \\ \mu + \alpha_1 + e_{12} \\ \mu + \alpha_1 + e_{13} \\ \mu + \alpha_2 + e_{21} \\ \mu + \alpha_2 + e_{22} \\ \mu + \alpha_3 + e_{31} \\ \mu + \alpha_3 + e_{32} \end{bmatrix} \tag{24}$$

or

$$\begin{bmatrix} 74 \\ 68 \\ 77 \\ 76 \\ 80 \\ 85 \\ 93 \end{bmatrix} = \mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} = \mathbf{Xb} + \mathbf{e}.$$

Thus,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}, \tag{25}$$

with $\mathbf{y}$ being the vector of observations and $\mathbf{e}$ the corresponding error terms.

General formulation of the model (1) for the one-way classification is achieved by writing:

1. the vector of responses as

$$\mathbf{y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_1} & \cdots & y_{i1} & y_{i2} & \cdots & y_{in_i} & \cdots & y_{a1} & y_{a2} & \cdots & y_{an_a} \end{bmatrix}; \tag{26}$$

2. the vector of parameters as

$$\mathbf{b}' = \begin{bmatrix} \mu & \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix}. \tag{27}$$

As a result, the matrix $\mathbf{X}$ has order $N \times (a + 1)$, where

$$N = n_{\cdot} = \sum_{i=1}^{a} n_i.$$

The symbols $N$ and $n_{\cdot}$ are used interchangeably.

The form of $\mathbf{X}$ in (25) is typical of its general form. Its first column is $\mathbf{1}_{\mathrm{N}}$ and of its other columns, the $i$th one has $\mathbf{1}_{n_i}$ in its $\left(\sum_{k=1}^{i-1} n_k + 1\right)$ th to $\left(\sum_{k=1}^{i} n_k\right)$ th rows, and zeros elsewhere. Thus, in these $a$ columns, the $\mathbf{1}_{n_i}$ vectors lie down the "diagonal," as in (25), and so can be written as a direct sum using the following notation.

**Notation.** The direct sum of three matrices $\mathbf{A}_1, \mathbf{A}_2,$ and $\mathbf{A}_3$ is defined (e.g., Searle (1966), Section 8.9) as

$$\sum_{i=1}^{3+} \mathbf{A}_i = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix}.$$

The symbol $\Sigma^+$ for a direct sum is introduced here for subsequent convenience. Using $\Sigma^+$, the form of $\mathbf{X}$ in the general one-way classification is, as in (25)

$$\mathbf{X} = \left[ \mathbf{1}_N \quad \sum_{i=1}^{a} {}^{+}\mathbf{1}_{n_i} \right]. \tag{28}$$

### b. The Normal Equations

The normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ of (2) are, from (26) and (28),

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \begin{bmatrix} n. & n_1 & n_2 & n_3 & \cdots & n_a \\ n_1 & n_1 & 0 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & 0 & \cdots & 0 \\ n_3 & 0 & 0 & n_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ n_a & 0 & 0 & 0 & \cdots & n_a \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \vdots \\ \alpha_a^\circ \end{bmatrix} = \begin{bmatrix} y.. \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ \vdots \\ y_{a.} \end{bmatrix} = \mathbf{X}'\mathbf{y}. \tag{29}$$

We see that $\mathbf{X}'\mathbf{X}$ has $n. = N$ as its leading element. The rest of the first row and column consists of the $n_i$'s. The $n_i$'s are also the remaining elements of the diagonal. The right-hand side of the equation, $\mathbf{X}'\mathbf{y}$, is the vector of response totals, that is, totals of the $y_{ij}$'s. The first is the grand total. The others are the class totals. For the data in Table 6.1, from (24) and (25), the normal equations are

$$\begin{bmatrix} 7 & 3 & 2 & 2 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \end{bmatrix} = \begin{bmatrix} y.. \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix} = \begin{bmatrix} 553 \\ 219 \\ 156 \\ 178 \end{bmatrix}. \tag{30}$$

The normal equations above clearly have the form of (29), with the right-hand vector $\mathbf{X}'\mathbf{y}$ having as elements the totals shown in Table 6.1.

### c. Solving the Normal Equations

Solving the normal equations (29) by means of (4) demands ascertaining the rank of $\mathbf{X}$ (or equivalently of $\mathbf{X}'\mathbf{X}$). In both (25) and (28), it is clear that the first column equals the sum of the others. This is also the case for $\mathbf{X}'\mathbf{X}$ of (29) and (30). The order of $\mathbf{X}$ is $p = a + 1$. Then the rank of $\mathbf{X}$, $r(\mathbf{X}) = a + 1 - 1 = a$. Thus, $p - r = 1$. Hence, by (4), we can solve the normal equations by putting one element of $\mathbf{b}^\circ$ equal to zero,

and crossing out one equation. In (29) and (30), we equate $\boldsymbol{\mu}^\circ$ to zero and delete the first equation. As a result, a solution to (29) is

$$\mathbf{b}^\circ = \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \vdots \\ \alpha_a^\circ \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{y}_{1.} \\ \bar{y}_{2.} \\ \vdots \\ \bar{y}_{a.} \end{bmatrix}. \tag{31}$$

Thus, a set of solutions to the normal equations is $\mu^\circ = 0$ and $\alpha_i^\circ = \bar{y}_{i.}$ for $i = 1, 2, \ldots, a$. The corresponding generalized inverse of $\mathbf{X}'\mathbf{X}$ is

$$\mathbf{G} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}\left\{\frac{1}{n_i}\right\} \end{bmatrix}, \tag{32}$$

where $\mathbf{D}\left\{\frac{1}{n_i}\right\}$ is the diagonal matrix of elements $1/n_i$ for $i = 1, 2, \ldots, a$. Multiplying $\mathbf{G}$ of (32) and $\mathbf{X}'\mathbf{X}$ of (29), we see that

$$\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{1}_a & \mathbf{I}_a \end{bmatrix}. \tag{33}$$

For the numerical example $\mathbf{b}^\circ$ of (31) is

$$\mathbf{b}^\circ = \begin{bmatrix} 0 \\ \frac{219}{3} \\ \frac{156}{2} \\ \frac{178}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 73 \\ 78 \\ 89 \end{bmatrix}. \tag{34}$$

From (32),

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \tag{35}$$

and

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 7 & 3 & 2 & 2 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{36}$$

### d. Analysis of Variance

In all cases, it is easy to compute SSM of (7) and SST of (9). The other term basic to the analysis of variance is SSR of (8). From (29) and (31), this is

$$\text{SSR} = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{a} \bar{y}_{i.} y_{i.} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i}. \tag{37}$$

For the data of Table 6.1, the calculation of these terms proceeds as follows. First,

$$\text{SSM} = N\bar{y}^2 = N\left(\frac{y_{..}}{N}\right)^2 = \frac{y_{..}^2}{N} = \frac{(553)^2}{7} = 43,687. \tag{38}$$

Second,

$$\text{SST} = \sum y^2 = 74 + 68^2 + 77^2 + 76^2 + 80^2 + 85^2 + 93^2 = 44,079. \tag{39}$$

Third, from (37),

$$\text{SSR} = \frac{219^2}{3} + \frac{156^2}{2} + \frac{178^2}{2} = 43,997. \tag{40}$$

Hence from (10),

$$\begin{aligned} \text{SSE} &= \text{SST} - \text{SSR} \\ &= \sum y^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} = 44,079 - 43,997 = 82. \end{aligned}$$

From (11) and (12),

$$\text{SSR}_{\text{m}} = \text{SSR} - \text{SSM} = 43,997 - 43,687 = 310$$

**TABLE 6.2 Analysis of Variance of Data in Table 6.1**

| Term | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Model (after mean) | $a - 1 = 2$ | $SSR_m = 310$ | $MSR_m = 155$ | $F(R_m) = 7.56$ |
| Residual error | $N - a = 4$ | $SSE = 82$ | $MSE = 20.5$ | |
| Total (after mean) | $N - 1 = 6$ | $SST_m = 392$ | | |

and

$$SST_m = SST - SSM = 44,079 - 43,687 = 392.$$

Using these values, we formulate the Table 6.2 that shows the analysis of variance. It is based on Table 5.6b of Chapter 5. From this, the estimated error variance is

$$\hat{\sigma}^2 = MSE = 20.5. \tag{41}$$

The coefficient of determination, as in (13), is

$$R^2 = \frac{SSR_m}{SST_m} = \frac{310}{392} = 0.79.$$

Thus, fitting the model $y_{ij} = \mu + \alpha_i + e_{ij}$ accounts for 79% of the total sum of squares.

The statistic $F(R_m)$ of (14) is

$$F(R_m) = \frac{MSR_m}{MSE} = \frac{155}{20.5} = 7.56$$

with $r - 1 = 2$ and $N - r = 4$ degrees of freedom. On the basis of normality, comparison of this with the tabulated values of the $F_{2,6}$-distribution provides a test of whether the model over and above a mean accounts for statistically significant variation in $y$. Since the 5% critical value is 6.94 which is exceeded by $F(R_m) = 7.56$, we conclude that the model accounts for statistically significant variation in $y$. Using a TI83 calculator, we find that the $p$-value is 0.0438. Thus, while this $F$-statistic is significant at $\alpha = .05$, it is not significant at $\alpha = .01$. Similarly, calculating (15) from (38) and (41) gives

$$F(M) = \frac{43,687}{20.5} = 2131.1.$$

Since the 5% critical point of the $F_{1,4}$-distribution is 7.71, we reject the hypothesis $H$: $E(\bar{y}) = 0$. This can also be construed as rejecting the hypothesis $H$: $\mu = 0$ when ignoring the $\alpha$'s. Actually, based on the large size of the $F$-statistic, we could reject the hypothesis $H$: $E(\bar{y}) = 0$ at any reasonable level of significance.

Computer outputs and programs in R and SAS to obtain the information in Table 6.2 are given below.

R Program and output:

```
> index=c(74,68,77,76,80,85,93)
> edu=c(rep("hsi",3),rep("hsg",2),rep("cg",2))
> result=data.frame(index,edu)
> result
  index edu
1    74 hsi
2    68 hsi
3    77 hsi
4    76 hsg
5    80 hsg
6    85  cg
7    93  cg
> res=aov(index~edu,data=result)
> summary(res)
             Df  Sum Sq  Mean Sq   F value   Pr(>F)
edu          2    310      155.0      7.561     0.0438 *
Residuals    4     82      20.5
—
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SAS program and output:

```
>
data investment;
input index educationlevel;
cards;
74 1
68 1
77 1
76 2
80 2
85 3
93 3
proc glm;
class educationlevel;
model index=educationlevel;
run;
```

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **educationlevel** | 3 | 1 2 3 |
| **Number of Observations Read** | | 7 |
| **Number of Observations Used** | | 7 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: index**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 2 | 310.0000000 | 155.0000000 | 7.56 | 0.0438 |
| **Error** | 4 | 82.0000000 | 20.500000 | | |
| **Corrected Total** | 6 | 392.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **index Mean** | |
| 0.790816 | 5.731256 | | 4.527693 | 79.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **Education level** | 2 | 310.0000000 | 155.0000000 | 7.56 | 0.0438 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **Education level** | 2 | 310.0000000 | 155.0000000 | 7.56 | 0.0438 |

### e. Estimable Functions

The expected value of any observation is estimable. Thus $\mu + \alpha_i$ is estimable. Correspondingly, the b.l.u.e. of $\mu + \alpha_i$ is $\mu^\circ + \alpha_i^\circ$.

We use $\hat{}$ over an expression to denote the b.l.u.e. of that expression. Noting the values of $\mu^\circ$ and $\alpha_i^\circ$ from (31) gives

$$\widehat{\mu + \alpha_i} = \mu^\circ + \alpha_i^\circ = \bar{y}_{i.} \tag{42}$$

The variance of the b.l.u.e. of an estimable function is obtained by expressing that function as $\mathbf{q}'\mathbf{b}$. Its b.l.u.e. is then $\mathbf{q}'\mathbf{b}^\circ$. The variance of the b.l.u.e. is $\mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2$. For example, with $\mathbf{b}'$ of (27)

$$\mu + \alpha_1 = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \mathbf{b}.$$

Then with

$$\mathbf{q}' = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{43}$$

$\widehat{\mu + \alpha_1}$ of (42) is $\mathbf{q}'\mathbf{b}^\circ$. Thus,

$$\widehat{\mu + \alpha_i} = \bar{y}_{i.} = \mathbf{q}'\mathbf{b}^\circ.$$

Hence,

$$v(\widehat{\mu + \alpha_1}) = v(\bar{y}_{1.}) = \frac{\sigma^2}{n_1} = \mathbf{q}'\mathbf{G}\mathbf{q}. \tag{44}$$

Using $\mathbf{G}$ of (37) it is easy to verify that $\mathbf{q}'\mathbf{G}\mathbf{q} = 1/n_1$ for $\mathbf{q}'$ of (42), using $\mathbf{G}$ of (32).

The basic result about estimable functions is (42). It provides b.l.u.e.'s of all of the estimable functions. Any linear combination of the $\mu + \alpha_i$ is estimable. Its b.l.u.e. is the same linear combination of the $\widehat{\mu + \alpha_i}$, that is, of the $\bar{y}_{i.}$. Thus for scalars $\lambda_i$,

$$\sum_{i=1}^{a} \lambda_i(\widehat{\mu + \alpha_i}) \text{ is estimable, with b .l.u.e. } \sum_{i=1}^{a} \lambda_i \bar{y}_{i.}. \tag{45}$$

Equivalently,

$$\widehat{\sum_{i=1}^{a} \lambda_i(\mu + \alpha_i)} = \sum_{i=1}^{a} \lambda_i(\widehat{\mu + \alpha_i}) = \sum_{i=1}^{a} \lambda_i \bar{y}_{i.}. \tag{46}$$

Although the variance of this b.l.u.e. can be obtained as $\mathbf{q'Gq}\sigma^2$ by expressing the estimable function as $\mathbf{q'b}$, it follows from (46) that the variance depends solely on the variances and covariances of $\bar{y}_{i.}$. These are

$$v(\bar{y}_{i.}) = \frac{\sigma^2}{n_i} \text{ and } \operatorname{cov}(\bar{y}_{i.}, \bar{y}_{k.}) = 0 \text{ for } i \neq k.$$

Thus, from (46),

$$v\left[\widehat{\sum_{i=1}^{a} \lambda_i(\mu + \alpha_i)}\right] = v\left(\sum_{i=1}^{a} \lambda_i \bar{y}_{i.}\right) = \left(\sum_{i=1}^{a} \frac{\lambda_i^2}{n_i}\right)\sigma^2. \tag{47}$$

From (47), the $100(1-\alpha)\,\%$ symmetric confidence interval $\sum_{i=1}^{a} q_i(\mu + \alpha_i)$ is from (19),

$$\sum_{i=1}^{a} q_i(\widehat{\mu + \alpha_i}) \pm \hat{\sigma} t_{N-r,\frac{1}{2}} \sqrt{\sum_{i=1}^{a} \frac{q_i^2}{n_i}} = \sum_{i=1}^{a} q_i \bar{y}_{i.} \pm \hat{\sigma} t_{N-r,\frac{1}{2}} \sqrt{\sum_{i=1}^{a} \frac{q_i^2}{n_i}} \tag{48}$$

For example,

$$\widehat{\alpha_1 - \alpha_2} = (\widehat{\mu + \alpha_1}) - (\widehat{\mu + \alpha_2}) \tag{49}$$

is estimable, with $\lambda_1 = 1, \lambda_2 = -1, \text{and} \lambda_3 = 0$ in (45). Hence using (34) in (36),

$$\widehat{\alpha_1 - \alpha_2} = (\widehat{\mu + \alpha_1}) - (\widehat{\mu + \alpha_2}) = \bar{y}_{1.} - \bar{y}_{2.} = 73 - 78 = -5.$$

From (47), the variance of

$$\widehat{\alpha_1 - \alpha_2}$$

is

$$v(\widehat{\alpha_1 - \alpha_2}) = \left[\frac{1^2}{3} + \frac{(-1)^2}{2}\right]\sigma^2 = \frac{5}{6}\sigma^2.$$

From (48), the $100(1-\alpha)\%$ symmetric confidence interval on $\alpha_1 - \alpha_2$ is

$$-5 \pm \hat{\sigma} t_{4,\frac{1}{2}\alpha} \sqrt{\frac{5}{6}}.$$

Then, a 95% symmetric confidence interval on $\alpha_1 - \alpha_2$ is

$$-5 \pm 4.528(2.78)\sqrt{\frac{5}{6}}$$

or

$$(-16.491, 6.491).$$

Since this confidence interval contains zero, we fail to reject the hypothesis $H$: $\alpha_1 = \alpha_2$ at $\alpha = .05$. There does not seem to be a statistically significant difference in the investment indices of people who have and have not completed high school on the basis of these data. On the other hand, a 95% confidence interval on $\alpha_1 - \alpha_3$ would be

$$-16 \pm 4.528(2.78)\sqrt{\frac{5}{6}}$$

or

$$(-27.532, -4.468).$$

Thus, we would reject the hypothesis $H$: $\alpha_1 = \alpha_3$ at $\alpha = .05$. We would conclude that there is a statistically significant difference between the investment indices of people who have not completed high school and people who were college graduates.

To give another example, we observe that

$$3\alpha_1 + 2\alpha_2 - 5\alpha_3 = 3(\mu + \alpha_1) + 2(\mu + \alpha_2) - 5(\mu + \alpha_3) \tag{50}$$

is estimable with $\lambda_1 = 3, \lambda_2 = 2,$ and $\lambda_3 = -5$. Thus,

$$\begin{aligned} \widehat{3\alpha_1 + 2\alpha_2 - 5\alpha_3} &= 3(\widehat{\mu + \alpha_1}) + 2(\widehat{\mu + \alpha_2}) - 5(\widehat{\mu + \alpha_3}) \\ &= 3(73) + 2(78) - 5(89) \\ &= -70. \end{aligned}$$

From (47),

$$\widehat{v(3\alpha_1 + 2\alpha_2 - 5\alpha_3)} = \left(\frac{3^2}{3} + \frac{2^2}{2} + \frac{5^2}{2}\right)\sigma^2 = 17.5\sigma^2$$

Again, using $\hat{\sigma}^2 = 20.5$ the $100(1-\alpha)\%$ on $3\alpha_1 + 2\alpha_2 - 5\alpha_3$ is

$$-70 \pm \sqrt{20.5} t_{4,\frac{1}{2}\alpha} \sqrt{17.5}$$

A 95% confidence interval would be

$$-70 \pm 4.528(2.78)\sqrt{17.5}$$

or

$$(-122.659, -17.341).$$

Thus at $\alpha = .05$, we would reject $H$: $3\alpha_1 + 2\alpha_2 - 5\alpha_3 = 0$.

Certain implications of (45) are worth noting. To be able to better explain these implications, we rewrite it in a slightly different but equivalent form. Observe that

$$\sum_{i=1}^{a} \lambda_i(\mu + \alpha_i) = \mu \sum_{i=1}^{a} \lambda_i + \sum_{i=1}^{a} \lambda_i \alpha_i \text{ is estimable.} \tag{51}$$

Observe that $r(\mathbf{X}) = a$. Thus, from Section 4f of Chapter 5, the maximum number of LIN estimable functions is $a$. Since there are $a$ functions $\mu + \alpha_i$, that are estimable, they constitute a LIN set of estimable functions. Hence, all other estimable functions are linear combinations of the $\mu + \alpha_i$. They are of the form (51). In more formal mathematical language, we can say that the $a$ estimable functions $\mu + \alpha_i$ constitute a basis for the vector space of estimable functions of dimension $a$. Some very important results about estimability for the one-way classification follow from this. They are presented in Theorem 1 below.

**Theorem 1** The following are properties of some estimable functions.

(i) The individual function $\mu$ is not estimable.
(ii) The individual functions $\alpha_i$ are not estimable.
(iii) The linear function $\left(\sum_{i=1}^{a} \lambda_i\right) \mu + \sum_{i=1}^{a} \lambda_i \alpha_i$ is estimable.
(iv) The linear combination $\sum_{i=1}^{a} \lambda_i \alpha_i$, where $\sum_{i=1}^{a} \lambda_i = 0$ is estimable.
(v) The differences $\alpha_i - \alpha_k$ are estimable for every $i \neq k$.

*Proof.*

(i) Suppose that $\mu$ is estimable. Then for some set of $\lambda_i$ values, (51) must reduce to $\mu$. For these $\lambda_i$, we would then have

$$\mu = \mu \sum_{i=1}^{a} \lambda_i + \sum_{i=1}^{a} \lambda_i \alpha_i \text{ identically.}$$

For this to hold true, the $\lambda_i$ must satisfy two conditions. They are

$$\sum_{i=1}^{a} \lambda_i = 1 \quad \text{and} \quad \sum_{i=1}^{a} \lambda_i \alpha_i = 0 \quad \text{for all} \quad \alpha_i.$$

The second of these conditions can only be true if for all $i$, $\lambda_i = 0$. Then $\sum_{i=1}^{a} \lambda_i \neq 1$, so the first condition is not true. Hence no $\lambda_i$ exist such that (51) reduces to $\mu$. Thus $\mu$ is not estimable.

(ii) Suppose $\alpha_k$ is estimable for some subscript $k$. Then in the second term of (51), we must have $\lambda_k = 1$ and $\lambda_i = 0$ for all $i \neq k$. Then (51) becomes $\mu + \alpha_k$. Hence $\alpha_k$ is not estimable.

(iii) This is simply a restatement of (51). It is made for purposes of emphasizing the estimability of any linear combination of $\mu$ and the $\alpha_i$'s in which the coefficient of $\mu$ is the sum of the coefficients of the $\alpha_i$. From (46), its b.l.u.e. is

$$\widehat{\left(\sum_{i=1}^{a} \lambda_i\right)\mu + \sum_{i=1}^{a} \lambda_i \alpha_i} = \sum_{i=1}^{a} \lambda_i \bar{y}_{i.}$$

For example, $13.7\mu + 6.8\alpha_1 + 2.3\alpha_2 + 4.6\alpha_3$ is estimable and its b.l.u.e. is $6.8\bar{y}_{1.} + 2.3\bar{y}_{2.} + 4.6\bar{y}_{3.}$. Two other estimable functions of more likely interest are

$$\mu + \frac{1}{N}\sum_{i=1}^{a} n_i \alpha_i \quad \text{with b.l.u.e.} \bar{y}_{..} \tag{52}$$

and

$$\mu + \frac{1}{a}\sum_{i=1}^{a} \alpha_i \quad \text{with b.l.u.e.} \sum_{i=1}^{a} \frac{\bar{y}_{i.}}{a}. \tag{53}$$

These are (45) – or, equivalently (51) – and (46) with $\lambda_i = n_i/n$ in (52) and $\lambda_i = 1/a$ in (53). For balanced data, $n_i = n$ for all $i$ and then (52) and (53) are the same.

(iv) This is just a special case of (51), where $\sum_{i=1}^{a} \lambda_i = 0$. It is (51) with $\mu$ eliminated. This shows that that any linear combination of the $\alpha_i$'s where the sum of the coefficients is zero is estimable. From (46), its b.l.u.e. is

$$\widehat{\sum_{i=1}^{\alpha} \lambda_i \alpha_i} \quad \text{with} \quad \sum_{i=1}^{a} \lambda_i = 0. \tag{54}$$

An example of an estimable function of the type in (54) is $3.6\alpha_1 + 2.7\alpha_2 - 6.3\alpha_3$ with b.l.u.e. $3.6\bar{y}_{1.} + 2.7\bar{y}_{2.} - 6.3\bar{y}_{3.}$. Another example is $\alpha_1 + \alpha_2 - 2\alpha_3$ or $\frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2 - \alpha_3$.

(v) This arises as a special case of the result in (iv). Putting $\lambda_i = 1$ and $\lambda_k = -1$ and all other $\lambda'$s zero shows that

$$\widehat{\alpha_i - \alpha_k} \text{ is estimable for every } i \neq k. \tag{55}$$

The difference between any pair of $\alpha$'s is estimable. By (46), its estimator is

$$\widehat{\alpha_i - \alpha_k} = \bar{y}_{i.} - \bar{y}_{k.}.$$

The variance of these differences is

$$\operatorname{var}(\widehat{\alpha_i - \alpha_k}) = \left(\frac{1}{n_i} + \frac{1}{n_k}\right)\sigma^2.$$

A $100(1-\alpha)\,\%$ symmetric confidence interval on $\alpha_i - \alpha_k$ is

$$\bar{y}_{i.} - \bar{y}_{k} \pm t_{N-r,\frac{1}{2}\alpha}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}.$$

The differences $\alpha_i - \alpha_k$ are frequently called *contrasts* (see subsection g that follows). All linear combinations of these differences are often called contrasts. They are estimable in accord with the principles of (46), (47), and (48). For example,

$$\alpha_1 + \alpha_2 - 2\alpha_3 = (\alpha_1 - \alpha_3) + (\alpha_2 - \alpha_3)$$

is estimable. ■

Of course, estimability of the above functions could be established from the basic property common to all estimable functions, that they are functions of expected values of observations. For example,

$$\alpha_1 - \alpha_2 = E(y_{1j}) - E(y_{2j}) = (\mu + \alpha_1) - (\mu + \alpha_2).$$

However, the detailed derivations show how particular cases are all part of the general result (42) to which all estimable functions belong.

### f. Tests of Linear Hypotheses

(*i*) ***General Hypotheses.*** The only hypotheses that can be tested are those that involve estimable functions. In all cases, they are tested using the statistic given by (21) of Section 1. We give an example using the data of Table 6.1. For these data, we have,

$$\mathbf{b}^{\circ\prime} = [\,0 \quad 73 \quad 78 \quad 89\,].$$

We consider the hypothesis

$$H: \quad \begin{array}{l} \alpha_2 - \alpha_1 = 9 \\ 2\alpha_3 - \alpha_1 - \alpha_2 = 30. \end{array}$$

This is equivalent to

$$\begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & -1 & 2 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 9 \\ 30 \end{bmatrix}.$$

We test this hypothesis by using

$$\mathbf{K}' = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & -1 & 2 \end{bmatrix}, \mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} 5 \\ 27 \end{bmatrix}$$

and

$$(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{17}{6} \end{bmatrix}^{-1} = \frac{1}{14} \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix}. \tag{56}$$

We have that

$$\mathbf{K}'\mathbf{b}^\circ - \mathbf{m} = \begin{bmatrix} 5 \\ 27 \end{bmatrix} - \begin{bmatrix} 9 \\ 30 \end{bmatrix} = \begin{bmatrix} -4 \\ -3 \end{bmatrix}.$$

As a result, $Q$ of (21) is

$$Q = \begin{bmatrix} -4 & -3 \end{bmatrix} \frac{1}{14} \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} -4 \\ -3 \end{bmatrix} = 24.357.$$

Using $s = r(\mathbf{K}') = 2$ and $\hat{\sigma}^2 = 20.5$ of (41),

$$F(H) = \frac{24.357}{2(20.5)} = .594 < 6.94$$

The value 6.94 is the upper 5% value of the $F_{2,4}$-distribution. Hence, we fail to reject the hypothesis.

***(ii) The Test Based on F(M).*** We test the hypothesis $H : E(\bar{\mathrm{y}}) = 0$ by using F(M) of (15).

Since, from the model (23), $NE(\bar{y}) = N\mu + \sum_{i=1}^{a} n_i\alpha_i$, this hypothesis is identical to

$H$: $N\mu + \sum_{i=1}^{a} n_i\alpha_i = 0$. To see that this is a testable hypothesis, we rewrite it as

$$H\text{: } \lambda'\mathbf{b} = \mathbf{0} \text{ with } \lambda' = \begin{bmatrix} N & n_2 & n_2 & \cdots & n_a \end{bmatrix}. \tag{57}$$

From (52), $\lambda'\mathbf{b}$ is estimable. To show that (21) reduces to SSM for (57), we use (31) and (32) to derive

$$\lambda'\mathbf{b} = \sum_{i=1}^{a} n_i\bar{y}_i = N\bar{y}.., \lambda'\mathbf{G} = \begin{bmatrix} 0 & \mathbf{1}' \end{bmatrix}, \text{ and } \lambda'\mathbf{G}\lambda = \sum_{i=1}^{a} n_i = N.$$

Hence, the numerator sum of squares for testing $H$ is, from (21),

$$Q = \mathbf{b}^{\circ\prime}\lambda(\lambda'\mathbf{G}\lambda)^{-1}\lambda'\mathbf{b}^{\circ} = N\bar{y}_{\cdot\cdot}^2 = \text{SSM}.$$

Furthermore, $s$ in (21) is defined as $s = r(\mathbf{K}')$. Here $s = r(\lambda') = 1$. Thus (21) is

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{\text{SSM}}{\hat{\sigma}^2} = \frac{\text{SSM}}{\text{SSE}} = F(M)$$

of (15). Hence, the $F$-test using $F(M)$ does test $H$: $N\mu + \sum_{i=1}^{a} n_i\alpha_i = 0$ or equivalently

$$\mu + \sum_{i=1}^{a} n_i\alpha_i/N = 0.$$

**Example 1 Testing $H$: $7\mu+3\alpha_1+2\alpha_2+2\alpha_3= 0$ for the Data of Table 6.2** We have that

$$\lambda'\mathbf{b}^{\circ} = \begin{bmatrix} 7 & 3 & 2 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ 73 \\ 78 \\ 89 \end{bmatrix} = 553 \text{ and } \lambda'\mathbf{G}\lambda = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 7 \\ 3 \\ 2 \\ 2 \end{bmatrix} = 7.$$

As a result,

$$Q = \frac{(553)^2}{7} = 43,687 = \text{SSM}$$

of (38). Hence,

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{43,687}{20.5} = 2131.1 = F(M)$$

as was calculated earlier. □

***(iii) The Test Based on $F(R_m)$.*** The test based on $F(R_m)$ shown in (14) is equivalent (for the one-way classification) to testing

$H$: all $\alpha'$s equal.

This in turn is equivalent to testing that all the $\alpha$'s are zero.

**Example 2 Test of Hypothesis of Equality of the $\alpha$'s for Data of Table 6.2** For the model in (24), there are only three $\alpha$'s. The above hypothesis $H$: $\alpha_1 = \alpha_2 = \alpha_3$ is identical to $H$: $\alpha_1 - \alpha_2 = \alpha_1 - \alpha_3 = 0$. This can be written as

$$H: \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{58}$$

Writing this as $\mathbf{K}'\mathbf{b} = \mathbf{0}$ we have,

$$\mathbf{K}' = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} -5 \\ 16 \end{bmatrix} \text{ and } (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \begin{bmatrix} \frac{5}{6} & \frac{1}{3} \\ \frac{1}{3} & \frac{5}{6} \end{bmatrix}^{-1}$$
$$= \frac{1}{7} \begin{bmatrix} 10 & -4 \\ -4 & 10 \end{bmatrix}$$

using $\mathbf{b}^\circ$ and $\mathbf{G}$ of (34) and (35). Hence in (21), where $s = r(\mathbf{K}') = 2$,

$$Q = \begin{bmatrix} -5 & -16 \end{bmatrix} \frac{1}{7} \begin{bmatrix} 10 & -4 \\ -4 & 10 \end{bmatrix} \begin{bmatrix} -5 \\ -16 \end{bmatrix} = 310 = \mathrm{SSR_m}$$

of Table 6.2. Therefore,

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{310}{2(20.5)} = 7.56 = F(R_\mathrm{m})$$

of Table 6.2. □

We now generalize the result of Example 2. We can write the hypothesis of equality of all the $\alpha$'s as

$$H: \mathbf{K}'\mathbf{b} = \mathbf{0} \text{ with } \mathbf{K}' = \begin{bmatrix} 0\mathbf{1}_{a-1} & \mathbf{1}_{a-1} & -\mathbf{I}_{a-1} \end{bmatrix}, \tag{59}$$

where $\mathbf{K}'$ has full-row rank $s = a - 1$. We can then show that $Q$ of (21) reduces to

$$Q = \sum_{i=1}^{a} n_i \bar{y}_{i.}^2 - N\bar{y}^2 = \mathrm{SSR} - \mathrm{SSM} = \mathrm{SSR_m},$$

using SSR defined in (37). Thus,

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{\text{SSR}_\text{m}}{(a-1)\text{MSE}} = \frac{\text{MSR}_\text{m}}{\text{MSE}} = F(R_\text{m})$$

as illustrated above. (See Exercise 20, Chapter 7.) Thus, the test statistic $F(R_m)$ provides a test of the hypothesis $H$: all $\alpha'$s equal.

We now consider the apparent equivalence of the preceding hypothesis to one in which all the $\alpha_i$'s are zero. First, we note that because $\alpha_i$ is not estimable the hypothesis

$H$: $\alpha_i = 0$ cannot be tested. Therefore, $H$: all $\alpha_i'$s $= 0$ cannot, formally, be tested. However, we can show that there is an apparent equivalence of the two hypotheses. Consider $Q$, the numerator sum of squares for testing $H : \mathbf{K}'\mathbf{b} = \mathbf{0}$. The identity

$$Q = \text{SSR} - (\text{SSR} - \text{Q})$$

is from Tables 3.8 and 5.9 equivalent to

$Q =$ SSR $-$ sum of squares due to fitting the reduced model.

Now, for the one-way classification based on

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{60}$$

we have just seen that the hypothesis $H$: all $\alpha'$s equal can be expressed in the form $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$ and tested. To carry out this test, we derive the underlying reduced model by putting all $\alpha_i$'s equal (to $\alpha$ say) in (60), thus getting

$$y_{ij} = \mu + \alpha + e_{ij} = \mu' + e_{ij}$$

as the reduced model (with $\mu' = \mu + \alpha$). The sum of squares for fitting this model is the same as that for fitting $y_{ij} = \mu + e_{ij}$ derived from putting $\alpha_i = 0$ in (60). Thus, the reduced model for $H$: all $\alpha_i'$s equal appears indistinguishable from that for $H$: all $\alpha_i'$s zero. Hence, the test based on $F(R_\text{m})$ sometimes gets referred to as testing $H$: all $\alpha_i'$s zero. More correctly, it is testing $H$: all $\alpha_i'$s equal.

### g. Independent and Orthogonal Contrasts

The general form of a contrast among effects $\alpha_i$ is a linear combination

$$\sum_{i=1}^{a} k_i\alpha_i = \mathbf{k}'\mathbf{b} \quad \text{with} \quad \mathbf{k}' = \begin{bmatrix} 0 & k_1 & \cdots & k_a \end{bmatrix} \quad \text{and} \quad \sum_{i=1}^{a} k_i = 0. \tag{61}$$

All such contrasts are orthogonal to $N\mu + \sum_{i=1}^{a} n_i\alpha_i$ that was considered in (57) because (22) is then satisfied. We have that for $\lambda'$ of (57), $\mathbf{G}$ of (32) and $\sum_{i=1}^{a} k_i$ of (61),

$$\lambda'\mathbf{G}\mathbf{k} = \begin{bmatrix} 0 & \mathbf{1}' \end{bmatrix}\mathbf{k} = \sum_{i=1}^{a} k_i = 0. \tag{62}$$

Thus, as a result of (62), (22) is satisfied. Furthermore, when testing a hypothesis that $(a-1)$ LIN such contrasts are zero $Q = \mathrm{SSR_m}$. For example, when testing

$$H\colon \begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & -1 & 1 \end{bmatrix} \mathbf{b} = \mathbf{0},$$

the values in (56) give $Q$ of (21) as

$$Q = \begin{bmatrix} 5 & 27 \end{bmatrix} \frac{1}{14} \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} 5 \\ 27 \end{bmatrix} = 310 = \mathrm{SSR_m}$$

of Table 6.2. □

The simplest forms of contrasts are differences between pairs of the $\alpha_i$'s. Such differences are the basis of the hypotheses considered in (58) and (59), which also satisfy (62). Hence the numerators of $F(M)$ and $F(R_\mathrm{m})$ are independent—as already established in Section 3 of Chapter 5 for the general case.

**Example 3 Finding an Orthogonal Contrast** Although, for example, $\alpha_1 - \alpha_2$ and $\alpha_1 - \alpha_3$ are both orthogonal to $7\mu + 3\alpha_1 + 2\alpha_2 + 2\alpha_3$, they are not orthogonal to each other. To find a contrast $\sum_{i=1}^{3} k_i\alpha_i$ orthogonal to $\alpha_1 - \alpha_2$, it is necessary that (22) is satisfied for $\sum_{i=1}^{3} k_i\alpha_i$ and $\alpha_1 - \alpha_2$. That means that

$$\begin{bmatrix} 0 & k_1 & k_2 & k_3 \end{bmatrix} \mathbf{G} \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} = \frac{k_1}{3} - \frac{k_2}{2} = 0$$

and $\sum_{i=1}^{3} k_i = 0$

Any $k$'s of the form $k_1 = -0.6k_3, k_2 = -0.4k_3$, and $k_3$ will suffice. For example, $\mathbf{k}' = \begin{bmatrix} 0 & -3 & -2 & 5 \end{bmatrix}$ gives $\mathbf{k}'\mathbf{b} = -3\alpha_1 - 2\alpha_2 + 5\alpha_3$. This contrast is orthogonal to both $\alpha_1 - \alpha_2$ and $7\mu + 3\alpha_1 + 2\alpha_2 + 2\alpha_3$. Testing

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & -3 & -2 & 5 \end{bmatrix} \mathbf{b} = 0$$

then involves

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} -5 \\ 70 \end{bmatrix} \text{ and } (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \begin{bmatrix} \frac{5}{6} & 0 \\ 0 & \frac{35}{2} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{6}{5} & 0 \\ 0 & \frac{2}{35} \end{bmatrix}.$$

Thus, $Q$ of (21) is

$$Q = \begin{bmatrix} -5 & 70 \end{bmatrix} \begin{bmatrix} \frac{6}{5} & 0 \\ 0 & \frac{2}{35} \end{bmatrix} \begin{bmatrix} -5 \\ 70 \end{bmatrix} = 30 + 280 = 310 = \mathrm{SSR_m}.$$

□

We can verify that the terms that make up the sum in the above equation are the numerator sums of squares for testing $\alpha_1 - \alpha_2 = 0$ and $-3\alpha_1 - 2\alpha_2 + 5\alpha_3 = 0$, respectively.

The fact that the off-diagonal elements of $\mathbf{K'GK}$ are zero provides further evidence of the truth of the proceeding statement and shows that the elements of $\mathbf{K'b}^\circ$ are independent.

### h. Models that Include Restrictions

As was emphasized in Section 6 of Chapter 5, linear models do not need to include restrictions on their elements. However, if restrictions are included, estimable functions and testable hypotheses may take different forms from those they have in the unrestricted model. In particular, functions of interest that are not estimable in the unrestricted model may be estimable in the restricted model.

In considering restrictions, we limit ourselves to those relating to non-estimable functions. The reason we impose this limitation is that restrictions relating to estimable functions do not alter the form of estimable functions and testable hypotheses available in the unrestricted model. Table 5.13 shows this. We also see there that the only changes from an unrestricted model incurred by having a restricted model are those wrought in estimable functions by the restriction that is part of the restricted model. These are particularly interesting in the one-way classification. We now illustrate some of them in this context.

Suppose the restricted model has the restriction $\sum_{i=1}^{a} n_i\alpha_i = 0$. The function $\mu + \sum_{i=1}^{a} n_i\alpha_i/n_\cdot$, which is estimable in the unrestricted model (as in (52)), becomes $\mu$ in the restricted model. By (52), $\mu$ has b.l.u.e. $\bar{y}_{\cdot\cdot}$. Note that $\mu$ is not estimable in the unrestricted model. However, in the restricted model with the restriction $\sum_{i=1}^{a} n_i\alpha_i = 0$, $\mu$ is estimable with b.l.u.e. $\bar{y}_{\cdot\cdot}$. Furthermore, the hypothesis considered in (57) and tested by means of $F(M)$ then becomes $H$: $\mu = 0$. Thus, under the restriction $\sum_{i=1}^{a} n_i\alpha_i = 0$, the $F$-statistic $F(M)$ can be used to test the hypothesis $H$: $\mu = 0$.

Suppose the model included the restriction $\sum_{i=1}^{a} \alpha_i = 0$. In the unrestricted model, $\mu + \sum_{i=1}^{a} \alpha_i/a$ is estimable with b.l.u.e. $\sum_{i=1}^{a} \bar{y}_{i\cdot}/a$ as in (53). This means that in the restricted model with $\sum_{i=1}^{a} \alpha_i = 0$, $\mu$ is estimable with b.l.u.e. $\sum_{i=1}^{a} \bar{y}_{i\cdot}/a$. In this case, the hypothesis $H$: $\mu = 0$ is tested by the $F$-statistic derived in the unrestricted model for testing $H$: $\mu + \sum_{i=1}^{a} \alpha_i/a = 0$. This is $H$: $\mathbf{k'b} = 0$ for $\mathbf{k'} = \begin{bmatrix} 1 & a^{-1}\mathbf{1'} \end{bmatrix}$, for which $\mathbf{k'b}^\circ = \sum_{i=1}^{a} \bar{y}_{i\cdot}/a$ and $\mathbf{k'Gk} = a^{-2} \sum_{i=1}^{a} (1/n_i)$. Hence the $F$-statistic for testing $H$: $\mu = 0$ in this restricted model is

$$F(H) = \frac{\left(\sum_{i=1}^{a} \bar{y}_{i\cdot}\right)^2}{\hat{\sigma}^2 \sum_{i=1}^{a} \frac{1}{n_i}}.$$

The preceding two paragraphs illustrate how different restrictions can lead to the same parameter being estimable in different restricted models even though that parameter

may not be estimable in the unrestricted model. Furthermore, even though it is formally the same parameter in the different restricted models (i.e., the same symbol), its b.l.u.e. in those models may not be the same. Its b.l.u.e. is the b.l.u.e. of the estimable function in the unrestricted model from which the estimable function in the restricted model has been derived by application of the restriction. Thus in a model having $\sum_{i=1}^{a} n_i\alpha_i = 0$, the b.l.u.e of $\mu$ is $\bar{y}_{..}$, the b.l.u.e. of $\mu + \sum_{i=1}^{a} n_i\alpha_i/n.$ in the unrestricted model. However, in a model having $\sum_{i=1}^{a} \alpha_i = 0$, the b.l.u.e. of $\mu$ is $\sum_{i=1}^{a} \bar{y}_{i.}/a$, the b.l.u.e. of $\mu + \sum_{i=1}^{a} \alpha_i/a$ in the unrestricted model. As a third example, consider a model having $\sum_{i=1}^{a} w_i\alpha_i = 0$ for some weights $w_i$. In this case, $\mu$ is estimable with b.l.u.e. $\sum_{i=1}^{a} w_i\bar{y}_{i.}/\sum_{i=1}^{a} w_i$, this being the b.l.u.e. of $\mu + \sum_{i=1}^{a} w_i\alpha_i/\sum_{i=1}^{a} w_i$ in the unrestricted model. Here, the $F$-statistic for testing $H$: $\mu = 0$ comes from testing

$$H\colon \mathbf{k}'\mathbf{b} = 0 \text{ with } \mathbf{k}' = \begin{bmatrix} 1 & \frac{w_1}{w_.} & \cdots & \frac{w_a}{w_.} \end{bmatrix} \quad \text{for} \quad w_. = \sum_{i=1}^{a} w_i.$$

Thus,

$$\mathbf{k}'\mathbf{b}^\circ = \frac{\sum_{i=1}^{a} w_i\bar{y}_{i.}}{w_.} \quad \text{and} \quad \mathbf{k}'\mathbf{G}\mathbf{k} = \frac{\left(\sum_{i=1}^{a} w_i^2/n_i\right)}{w_.^2}.$$

As a result, the $F$-statistic for testing $H$: $\mu = 0$ is

$$F(H) = \frac{\left(\sum_{i=1}^{a} w_i\bar{y}_i\right)^2}{\hat{\sigma}^2 \sum_{i=1}^{a} \frac{w_i^2}{n_i}}.$$

Table 6.3 summarizes the three cases mentioned above. Of course, the first two rows of Table 6.3 are special cases of the last row. We have that $w_i = n_i$ for the first row and $w_i = 1$ for the second. In all three rows, $\mu$ is estimable. Since $\mu + \alpha_i$ is also estimable (anything estimable in the unrestricted model is estimable for the restricted model), it follows that in the restricted models, $\alpha_i$ is estimable with b.l.u.e. being $\bar{y}_{i.}$ minus the b.l.u.e. of $\mu$.

The choice of what model to use, the unrestricted model, one of those in Table 6.3 or some other depends on the nature of the data. For unbalanced data, we often find $\sum_{i=1}^{a} n_i\alpha_i = 0$ used. Having the same restrictions on the solutions $\sum_{i=1}^{a} n_i\alpha_i^\circ = 0$ leads to an easy procedure for solving the normal equations, as is evident from (29): $\mu^\circ = \bar{y}_{..}$ and $\alpha_i^\circ = \bar{y}_{i.} - \bar{y}_{...}$. This is perfectly permissible for finding a solution $\mathbf{b}^\circ$, it being of course, the oft-referred-to method of applying the "usual constraints" as discussed in Section 7 of Chapter 5. Although $\sum_{i=1}^{a} n_i\alpha_i^\circ = 0$ provides an easy solution for $\mathbf{b}^\circ$, the same restriction applied to the parameters $\sum_{i=1}^{a} n_i\alpha_i = 0$ may not always be appropriate. For example, suppose an experiment to estimate the efficacy of a feed additive for dairy cows is done on 7 Holsteins, 5 Jerseys, and 2 Guernseys. The "constraint" $7\alpha_1^\circ + 5\alpha_2^\circ + 2\alpha_3^\circ = 0$ would lead to solutions for $\mu^\circ, \alpha_1^\circ, \alpha_2^\circ,$ and $\alpha_3^\circ$ very

**TABLE 6.3 Estimators of $\mu$ and $F$-Statistics for Testing $H$: $\mu = 0$, in Three Different Restricted Models**

| Restriction on Model | Estimable Function in Unrestricted Model Which Reduces to $\mu$ in Restricted Model | b.l.u.e. of $\mu$ in Restricted Model = Blue of Function in Preceding Column in Unrestricted Model | $F$-Statistic for Testing $H$: $\mu = 0$ |
|---|---|---|---|
| $\sum_{i=1}^{a} n_i\alpha_i = 0$ | $\mu + \sum_{i=1}^{a} \frac{n_i\alpha_i}{n_\cdot}$ | $\bar{y}_{\cdot\cdot}$ | $F(M) = \frac{n\bar{y}_{\cdot\cdot}^2}{\hat{\sigma}^2}$ |
| $\sum_{i=1}^{a} \alpha_i = 0$ | $\mu + \sum_{i=1}^{a} \frac{\alpha_i}{a}$ | $\frac{\sum_{i=1}^{a} \bar{y}_{i\cdot}}{a}$ | $\frac{\left(\sum_{i=1}^{a} \bar{y}_{i\cdot}\right)^2}{\hat{\sigma}^2 \sum_{i=1}^{a} \frac{1}{n_i}}$ |
| $\sum_{i=1}^{a} w_i\alpha_i = 0$ | $\mu + \sum_{i=1}^{a} \frac{w_i\alpha_i}{w_\cdot}$ | $\frac{\sum_{i=1}^{a} w_i\bar{y}_{i\cdot}}{w_\cdot}$ | $\frac{\left(\sum_{i=1}^{a} w_i\bar{y}_{i\cdot}\right)^2}{\hat{\sigma}^2 \sum_{i=1}^{a} \frac{w_i^2}{n_i}}$ |

easily. However, if the proportion of these three breeds in the whole population of dairy cows (assumed to consist of these three breeds and no others) was 6:2:2, it would be more meaningful to use $6\alpha_1^\circ + 6\alpha_2^\circ + 2\alpha_3^\circ = 0$ rather than $7\alpha_1^\circ + 5\alpha_2^\circ + 2\alpha_3^\circ = 0$, if any such restriction was desired. In this case, we would use the third row of Table 6.3 rather than the first.

### i. Balanced Data

We now show how the results in the above discussion and Table 6.3 specialize for balanced data. With balanced data, $n_i = n$ for all $i$. Then the first two rows of Table 6.3 are the same. The "constraint" $\sum_{i=1}^{a} \alpha_i^\circ = 0$ provides an easy solution to the normal equations $\mu^\circ = \bar{y}_{\cdot\cdot}$ and $\alpha_i^\circ = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot\cdot}$. This solution is the one frequently found in the literature. Apart from this, all other results stand fast. For example, $\mu + \alpha_i$ and $\alpha_i - \alpha_k$ are estimable, with b.l.u.e.'s $\bar{y}_{i\cdot}$ and $\bar{y}_{i\cdot} - \bar{y}_{k\cdot}$, respectively. Furthermore, as usual, $\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{n} y_{i\cdot}^2/n$.

Sometimes, the restriction $\sum_{i=1}^{a} \alpha_i = 0$ is used as part of the model. This is in accord with the "constraint" $\sum_{i=1}^{a} \alpha_i^\circ = 0$, useful for solving the normal equations. As a restriction, it can also be opportunely rationalized in terms of defining the $\alpha_i$'s as deviations from their mean, and hence having their mean be zero, that is, $\sum_{i=1}^{a} \alpha_i = 0$. The effect of the restriction is to make $\mu$ and $\alpha_i$ estimable with b.l.u.e. $\hat{\mu} = \bar{y}_{\cdot\cdot}$ and $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$, and hypotheses about individual values of $\mu$ and $\alpha_i$ are then testable.

## 3. REDUCTIONS IN SUMS OF SQUARES

### a. The R( ) Notation

Consideration of models more complex than that for the one-way classification will lead us to comparing the adequacy of different models for the same set of data. Since in the identity SSE = SST − SSR, we have SSR as the reduction in the total sum of squares due to fitting any particular model, SSR is a measure of the variation in $y$ accounted for by that model. Comparing the values of SSR that result from fitting the different models can therefore make comparison of different models for a set of data. To facilitate discussion of these comparisons, we refer, as previously, to SSR as a reduction in sum of squares. We now denote it by $R(\ )$ with the contents of the brackets indicating the model fitted. For example, when fitting $y_{ij} = \mu + \alpha_i + e_{ij}$, the reduction in the sum of squares is $R(\mu, \alpha)$. This indicates a model that has parameters $\mu$ and those of an $\alpha$-factor. Likewise, $R(\mu, \alpha, \beta)$ is the reduction in the sum of squares for fitting $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$. Furthermore, $R(\mu, \alpha, \beta : \alpha)$ is the reduction due to fitting the nested model $y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$. The symbol $\beta : \alpha$ in $R(\mu, \alpha, \beta : \alpha)$ indicates that the $\beta$-factor is nested within the $\alpha$-factor. Extension to more complex models is straightforward. At all times, the letter $R$ is mnemonic for "reduction" in sum of squares and not for "residual" as used by some writers. In this book, $R(\ )$ is always a reduction in the sum of squares.

The model $y_i = \mu + e_i$ has normal equation $N\mu = y_.$. The corresponding reduction in the sum of squares, $R(\mu)$, is readily found to be $N\bar{y}^2$. However, for all models, $N\bar{y}^2$ is SSM. Therefore,

$$R(\mu) = N\bar{y}^2 = \text{SSM}.$$

For the one-way classification model, $y_{ij} = \mu + \alpha_i + e_{ij}$, the reduction in the sum of squares, now written as $R(\mu, \alpha)$ is, by (37)

$$\text{SSR} = R(\mu, \alpha) = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i}.$$

Therefore from (11),

$$\text{SSR}_\text{m} = \text{SSR} - \text{SSM} = R(\mu, \alpha) - R(\mu). \tag{63}$$

Thus for the one-way classification, $\text{SSR}_\text{m}$ is the difference between the reductions in the sums of squares due to fitting two different models, one containing $\mu$ and an $\alpha$-factor, and the other just containing $\mu$. Therefore, we can view $\text{SSR}_\text{m}$ of (63) as the additional reduction in the sum of squares due to fitting a model containing a $\mu$ and an $\alpha$-factor over and above fitting one just containing $\mu$. Hence $R(\mu, \alpha) - R(\mu)$ is the additional reduction due to fitting $\mu$ and $\alpha$ over and above fitting $\mu$. More succinctly, it is the reduction due to fitting $\alpha$ over and above $\mu$. An equivalent interpretation is that, once having fitted $\mu$, the difference $R(\mu, \alpha) - R(\mu)$ represents the reduction in the

sum of squares due to fitting an $\alpha$-factor additional to $\mu$. In this way, $R(\mu, \alpha) - R(\mu)$ is the reduction due to fitting "$\alpha$ after having already fitted $\mu$" or fitting "$\alpha$ after $\mu$." In view of this, we use the symbol $R(\alpha|\mu)$ for (63) and write

$$R(\alpha|\mu) = R(\mu, \alpha) - R(\mu). \tag{64}$$

It is easy to extend this notation. For example,

$$R(\alpha|\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta)$$

is the reduction in the sum of squares after fitting "$\alpha$ after $\mu$ and $\beta$." That means the reduction due to fitting a model containing $\mu$, an $\alpha$-factor and a $\beta$-factor. It is a measure of the extent that a model can explain more of the variation in $y$ by having in it, in a specified manner, something more than just $\mu$ and a $\beta$-factor.

Every $R(\ )$ term is by definition the SSR of some model. Therefore, its form is $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ for $\mathbf{X}$ appropriate to that model. The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is idempotent. Therefore for $y \sim N(\mu, \sigma^2\mathbf{I})$, for any vector $\boldsymbol{\mu}$, the distribution of $R()/\sigma^2$ is a non-central $\chi^2$ independent of SSE. Suppose $R(\mathbf{b}_1, \mathbf{b}_2)$ is the reduction for fitting $\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{Z}\mathbf{b}_2 + \mathbf{e}$ and $R(\mathbf{b}_1)$ is the reduction for fitting $\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{e}$. It can then be shown that (see Exercise 13) $R(\mathbf{b}_2|\mathbf{b}_1)/\sigma^2$ has a non-central $\chi^2$-distribution independent of $R(\mathbf{b}_1)$ and of SSE. Hence, whenever the reduction in the sum of squares $R(\mathbf{b}_1, \mathbf{b}_2)$ is partitioned as $R(\mathbf{b}_1, \mathbf{b}_2) = R(\mathbf{b}_2|\mathbf{b}_1) + R(\mathbf{b}_1)$, we know that both $R(\mathbf{b}_2|\mathbf{b}_1)$ and $R(\mathbf{b}_1)$ have non-central $\chi^2$-distributions and that they are independent of each other and of SSE.

The succinctness of the $R(\ )$ notation and its identifiability with its corresponding model is readily apparent. This, and the distributional properties just discussed, provide great convenience for considering the effectiveness of different models. As such, it is used extensively in what follows.

### b. Analyses of Variance

Table 6.2 is an example of the analysis of variance given in Table 5.6b of Chapter 5. Its underlying sums of squares can be expressed in terms of the $R(\ )$ notation as follows:

$$\begin{aligned} &\text{SSM} = R(\mu) = 43,687, \ \text{SSR} = R(\mu, \alpha) = 43,997, \\ &\text{SSR}_{\text{m}} = R(\alpha|\mu) = 310, \ \text{SSE} = \text{SST} - R(\mu, \alpha) = 82. \end{aligned}$$

These are summarized in Table 6.4. There, the aptness of the $R(\ )$ notation for highlighting the sum of squares is evident. We have that

1. The term $R(\mu)$ is the reduction due to fitting the mean $\mu$.
2. The term $R(\alpha|\mu)$ is the reduction due to fitting the $\alpha$-factor after $\mu$.
3. The term $R(\mu, \alpha)$ is the reduction due to fitting the model consisting of an $\alpha$-factor and $\mu$.

**TABLE 6.4 Analysis of Variance Using $R(\ )$ Notation. (See Also Tables 6.2 and 5.5)**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean | $1 = 1$ | $R(\mu) = 43{,}687$ | 43,687 | 2131.1 |
| $\alpha$-factor after mean | $a - 1 = 2$ | $R(\alpha\|\mu) = 310$ | 155 | 7.56 |
| Residual error | $N - \alpha = 4$ | $\text{SSE} = \text{SST} - R(\mu,\alpha) = 82$ | 20.5 | |
| Total | $N = 7$ | $\text{SST} = 44{,}079$ | | |

The attendant residual sum of squares is $\text{SSE} = \text{SST} - R(\mu, \alpha)$. Of course, as in (64) we have that,

$$R(\mu, \alpha) = R(\mu) + R(\alpha|\mu).$$

The clarity provided by the $R(\ )$ notation is even more evident for models that involve several factors. The notation is therefore used universally in all analysis variance tables that follow. Furthermore, all such tables have a format similar to that of Table 6.4 in that they show:

1. a line for the mean $R(\mu)$;
2. a total sum of squares $\text{SST} = \sum_{i=1}^{n} y_i^2$ not corrected for the mean.

The only quantity that such a table does not yield at a glance is the coefficient of determination $R^2 = \text{SSR}_\text{m}/\text{SST}_\text{m}$ of equation (13). However, since this can always be expressed as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST} - R(\mu)}, \tag{65}$$

it too can readily be derived from the analysis of variance tables such as Table 6.4.

### c. Tests of Hypotheses

In Section 2f(iii), we saw how $F(M)$ is a suitable statistic for testing $H$: $n.\mu + \sum_{i=1}^{a} n_i\alpha_i = 0$. However,

$$F(M) = \frac{\text{SSM}}{\text{MSE}} = \frac{R(\mu)}{\hat{\sigma}^2}.$$

This gives us a dual interpretation of $R(\mu)$. On the one hand, we see that it is the numerator sum of squares for testing $H$: $n.\mu + \sum_{i=1}^{a} n_i\alpha_i = 0$. On the other hand, it is the reduction in the sum of squares due to fitting the model $y_{ij} = \mu + e_{ij}$.

We also have a dual interpretation of $R(\alpha|\mu)$. As we have already mentioned, it is the reduction in the sum of squares due to fitting $\alpha$ after $\mu$. Section 2f(iii) explains why $F(R_{\mathrm{m}})$ is referred to as testing $H$: all $\alpha_i'$s equal. However,

$$F(R_{\mathrm{m}}) = \frac{\mathrm{MSR}_{\mathrm{m}}}{\mathrm{MSE}} = \frac{\mathrm{SSR}_{\mathrm{m}}}{(a-1)\mathrm{MSE}} = \frac{R(\alpha|\mu)}{(a-1)\hat{\sigma}^2}.$$

Thus, we see that $R(\alpha|\mu)$ is also the numerator sum of squares for testing $H$: all $\alpha_i'$s equal. The association of $R(\alpha|\mu) = R(\mu,\alpha) - R(\mu)$ with the effective testing of $H$: all $\alpha_i'$s equal is particularly convenient. Putting $\alpha = 0$ in the symbol $R(\mu,\alpha)$ reduces the symbol to $R(\mu)$. The difference between these two, $R(\mu,\alpha) - R(\mu)$, is the required numerator sum of squares.

In terms of $R(\alpha|\mu)$ being the numerator sum of squares for testing the hypothesis $H$: all $\alpha_i'$s equal, Table 6.4 is an application of Table 5.9. Writing $H$: all $\alpha_i'$s equal in the form $H : \mathbf{K}'\mathbf{b} = \mathbf{0}$ as in (59), we see in Table 5.9 that $R(\alpha|\mu)$ is the numerator sum of squares for testing $H : \mathbf{K}'\mathbf{b} = \mathbf{0}$, and $R(\mu)$ is the sum of squares for the reduced model $y_{ij} = \mu + \alpha + e_{ij} = \mu' + e_{ij}$.

## 4. MULTIPLE COMPARISONS

The results of an analysis of variance only tell us that some but not necessarily all of the effects differ from one another. We can perform $t$-tests on individual differences or linear combinations. However, if we have several such confidence intervals, the probability that each coordinate of a vector of means will lie in all of them will be greater than $1 - \alpha$. In Example 14 of Chapter 3, we gave one method of constructing a simultaneous confidence interval on two regression coefficients by halving the levels of significance for each of two individual confidence intervals. If we had $m$ such regression coefficients or linear combinations of the same, we could find individual $1 - \alpha/m$ confidence intervals. This is known as the Bonferroni method of finding simultaneous intervals.

Another method of finding simultaneous confidence intervals is due to Scheffe (see Scheffe (1959) or Hogg and Craig (2014)). We will present this method and give an outline of its derivation along the lines of Scheffe (1959).

Confidence sets are generalizations of confidence intervals. Suppose that $\{y_1, \ldots, y_n\}$ is a set of observations whose distribution is completely determined by the unknown values of the parameters $\{\theta_1, \ldots, \theta_m\}$ and that $\{\psi_1, \ldots, \psi_q\}$ are specified functions of the $\theta$ parameters. For example, in the context of linear models, the $\psi$'s could be estimable functions of the $\theta$ parameters. The set of all the $\psi$'s may be thought about as a $q$-dimensional space. Suppose that for every possible $\mathbf{y}$ in the sample space, we have a region $R(\mathbf{y})$ of the $q$-dimensional space. Suppose that $\psi$ is

thought of as a point determined by the value of $\theta$. If a region $R(\mathbf{y})$ has the property that the probability that it covers the true point $\psi$ is a pre-assigned constant $1 - \alpha$, no matter what the unknown true parameter point $\theta$ is, we say that $R(\mathbf{y})$ is a *confidence set* for $\psi$ with *confidence coefficient* $1 - \alpha$. If, for example, $\alpha = .05$ and we take 100 samples a very large number of times the proportion of confidence sets that actually contain the true point $\psi$ should average out to about 95. When $q = 1$ and $R(\mathbf{y})$ is an interval on the real line, the confidence set is a confidence interval.

Assume that $\psi_1, \psi_2, \ldots, \psi_q$ denote $q$ linearly independent estimable functions. Let $\hat{\psi}$ be an unbiased estimator of $\psi$. We have that $\psi = \mathbf{Cb}$ where the $\mathbf{b}$ are the parameters of the regression model and $\hat{\psi} = \mathbf{Ay}$. Scheffe finds a confidence set for $\{\psi_1, \ldots, \psi_q\}$ in a $q$-dimensional space. The confidence set takes the form of the ellipsoid

$$(\boldsymbol{\Psi} - \hat{\boldsymbol{\Psi}})' \mathbf{B} (\boldsymbol{\Psi} - \hat{\boldsymbol{\Psi}}) \leq q s^2 F_{\alpha,q,n-r} \tag{66}$$

Scheffe then states the following Theorem.

**Theorem 2** For a regression model, the probability is $1 - \alpha$ that simultaneously for all estimable functions $\psi$,

$$\hat{\psi} - S\hat{\sigma}_{\hat{\psi}} \leq \psi \leq \hat{\psi} + S\hat{\sigma}_{\hat{\psi}}, \tag{67}$$

where $S = (qF_{\alpha;q,n-r})^{1/2}$.

*Sketch of Proof.* The proof is based on the fact that for a point to lie inside the ellipsoid it must lie on a line connecting the points of contact with the ellipsoid of two parallel tangent planes. ■

**Example 4 Some Simultaneous Confidence Intervals** The data below were compiled by the National Center for Statistics and Analysis, United States. Some of the data points are missing. For five states in different parts of the country, the data represent the number of speeding-related fatalities by road type and speed limit in miles per hour during 2003 on non-interstate highways.

| State/Speed Limit | 55 | 50 | 45 | 40 | 35 | < 35 |
|---|---|---|---|---|---|---|
| California | 397 | 58 | 142 | 107 | 173 | – |
| Florida | – | – | – | – | 80 | 75 |
| Illinois | 226 | 3 | 22 | 47 | 69 | 88 |
| New York | 177 | 10 | 23 | 30 | 23 | – |
| Washington | 16 | – | 15 | 18 | 53 | 43 |

The results of a one-way ANOVA are below.

**The SAS System**

**The GLM Procedure**

**Dependent Variable: fatality**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 5 | 78240.9449 | 15648.1890 | 2.53 | 0.0694 |
| **Error** | 17 | 105300.5333 | 6194.1490 | | |
| **Corrected Total** | 22 | 183541.4783 | | | |

| R-Square | Coeff Var | Root MSE | fatality Mean |
|---|---|---|---|
| 0.426285 | 95.52333 | 78.70292 | 82.39130 |

| Level of speed | N | fatality | |
|---|---|---|---|
| | | Mean | Std. Dev. |
| **< 35** | 3 | 68.666667 | 23.158872 |
| **35** | 5 | 79.600000 | 56.451749 |
| **40** | 4 | 50.500000 | 39.501055 |
| **45** | 4 | 50.500000 | 61.103737 |
| **50** | 3 | 23.666667 | 29.938827 |
| **55** | 4 | 204.000000 | 156.850247 |

We shall find a 95% simultaneous confidence interval for the differences of adjacent means by both the Scheffe and Bonferonni method.

For the Scheffe method, the formula would be

$$\bar{y}_{i.} - \bar{y}_{j.} \pm \sqrt{(I-1)F_{\alpha,I-1,n-I}}\, s \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

We have that $I = 6, N - I = 17, s = 78.7029, F_{.05,5,17} = 2.81$

For Bonferonni, we would have

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_{\alpha/2(I-1)} s \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}.$$

For a 95% simultaneous confidence interval, we would use a $t$-value of 2.898.

< 35 versus 35

Scheffe

$$68.667 - 79.600 \pm \sqrt{5(2.81)\left(\tfrac{1}{3} + \tfrac{1}{5}\right)(6194.149)}$$
$$-10.933 \pm 215.44$$
$$(-226.37, 204.51)$$

Bonferonni

$$68.667 - 79.600 \pm 2.898\sqrt{\left(\tfrac{1}{3} + \tfrac{1}{5}\right)(6194.149)}$$
$$-10.933 \pm 166.567$$
$$(-177.5, 155.634)$$

35 versus 40

Scheffe

$$79.6 - 50.5 \pm \sqrt{5(2.81)\left(\tfrac{1}{5} + \tfrac{1}{4}\right)(6194.149)}$$
$$29.1 \pm 197.895$$
$$(-168.8, 227)$$

Bonferonni

$$79.6 - 50.5 \pm 2.898\sqrt{\left(\tfrac{1}{5} + \tfrac{1}{4}\right)(6194.149)}$$
$$29.1 \pm 153.001$$
$$(-123.901, 182.101)$$

40 versus 45

Scheffe

$$50.5 - 50.5 \pm \sqrt{5(2.81)\left(\tfrac{1}{4} + \tfrac{1}{4}\right)(6194.149)}$$
$$0 \pm 208.6$$
$$(-208.6, 208.6)$$

Bonferonni

$$50.5 - 50.5 \pm 2.898\sqrt{\left(\tfrac{1}{4} + \tfrac{1}{4}\right)(6194.149)}$$
$$0 \pm 161.278$$
$$(-161.278, 161.278)$$

45 versus 50

Sheffe

$$50.5 - 23.667 \pm \sqrt{5(2.81)\left(\tfrac{1}{4} + \tfrac{1}{3}\right)(6194.149)}$$
$$26.833 \pm 225.314$$
$$(-198.48, 252.15)$$

Bonferonni

$$50.5 - 23.667 \pm 2.898\sqrt{\left(\tfrac{1}{4} + \tfrac{1}{3}\right)(6194.149)}$$
$$26.833 \pm 174.2$$
$$(-147.367, 201.033)$$

50 versus 55

Scheffe

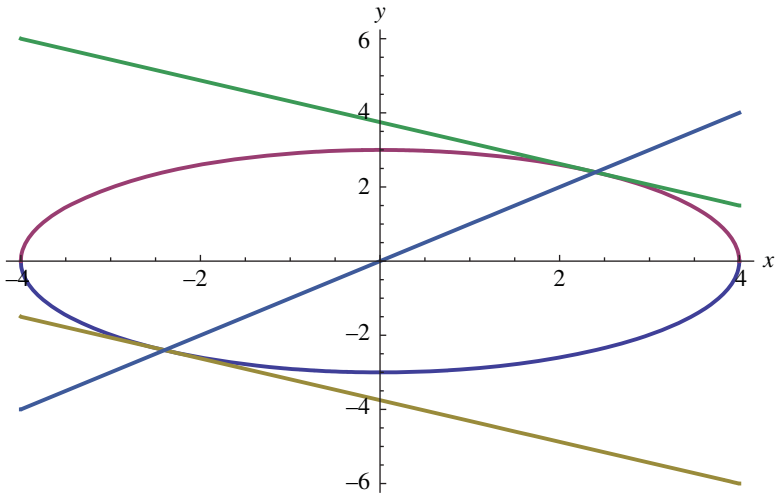
$$23.667 - 204 \pm \sqrt{5(2.81)\left(\tfrac{1}{4} + \tfrac{1}{3}\right)(6194.149)}$$
$$-180.33 \pm 225.314$$
$$(-405.64, 44.98)$$

Bonferonni

$$23.667 - 204 \pm 2.898\sqrt{\left(\tfrac{1}{4} + \tfrac{1}{3}\right)(6194.149)}$$
$$-180.33 \pm 174.2$$
$$(-354.33, -6.13)$$

Bonferonni intervals are generally narrower than those of Scheffe.

**Example 5 Illustration of Geometry of Scheffe Confidence Intervals** The point (1, 1) lies on the line connecting the points of contact of the tangent lines to the ellipse $\frac{x^2}{16} + \frac{y^2}{9} = 1$. The points of contact are (12/5, 12/5) and (−12/5, −12/5). The parallel tangent lines are $y = -(9/16)x - 15/4$ and $y = -(9/16)x + 15/4$. This illustrates in two dimensions, the principle used by Scheffe in deriving simultaneous confidence intervals, the fact that a point inside an ellipsoid or ellipse must lie on a line connecting points of contact of parallel tangent lines. □




For more information about multiple comparisons, see, for example, Hochberg and Tamhane (1987), Miller (1981), and Hsu (1996).

## 5. ROBUSTNESS OF ANALYSIS OF VARIANCE TO ASSUMPTIONS

Analysis of variance is performed assuming three basic assumptions are true. They are:

1. The errors are normally distributed.
2. The errors all have the same variance.
3. The errors are statistically independent.

The problem is how valid is the analysis when one or more of these assumptions are violated and what, if anything, can be done about it. We shall devote one subsection to the violation of each of the assumptions mentioned above.

### a. Non-normality of the Error

As Scheffe (1959) points out, two measures that may be used to consider the effects of non-normality are the measures $\gamma_1$ of skewness and $\gamma_2$ of kurtosis of a random variable $x$. Using the standard notation $\mu$ for the mean and $\sigma^2$ for the variance of a distribution, we may define the skewness as

$$\gamma_1 = \frac{1}{\sigma^3} E[(x - \mu)^3] \tag{68}$$

and kurtosis

$$\gamma_2 = \frac{1}{\sigma^4} E[(x-\mu)^4] - 3. \tag{69}$$

The skewness and kurtosis are indicators of the departure of the distribution of a random variable from normality. A positive skewness, that is, $\gamma_1 > 0$ indicates that on the right-hand side, the tail of the distribution appears to be flatter than on the left-hand side. Likewise, a negative skewness, that is, $\gamma_1 < 0$ indicates that the tail of the distribution is flatter on the left-hand side. However, for large samples, the skewness of functions of the sample data generally approaches zero because of the central limit theorem and inferences about means are not greatly affected by departures from normality.

However, as Scheffe (1959) explains in his discussion on page 336, inferences about variances can be affected a great deal by departures from normality. In particular, this can happen if the kurtosis $\gamma_2$ is significantly different from zero. Under normal theory, inferences about $\sigma^2$ are usually based on the distribution of $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. When the distribution is normal and the kurtosis is zero, we have that

$$E\left(\frac{s^2}{\sigma^2}\right) = 1, \; E\left(\frac{s^2}{\sigma^2}\right) = 1 \tag{70}$$

and

$$\text{var}\left(\frac{s^2}{\sigma^2}\right) = \frac{2}{n-1} \tag{71}$$

If the kurtosis differs from zero, we have,

$$\text{var}\left(\frac{s^2}{\sigma^2}\right) = \frac{2}{n-1} + \frac{\gamma_2}{n}. \tag{72}$$

Observe that the ratio of (72) to (71) is

$$1 + \frac{n-1}{2n}\gamma_2$$

and that

$$\lim_{n\to\infty}\left(1 + \frac{n-1}{2n}\gamma_2\right) = 1 + \frac{1}{2}\gamma_2.$$

Furthermore, by the central limit theorem, $s^2$ is normal. As a result,

$$\left(\frac{n-1}{2}\right)^{-1/2}\left(\frac{s^2}{\sigma^2} - 1\right) \sim N\left(0, 1 + \frac{1}{2}\gamma_2\right)$$

for large $n$ instead of $N(0, 1)$. This causes a serious error in any confidence level, significance level or power computation under normal theory when the kurtosis $\gamma_2$ is significantly different from zero. This will be of special importance in the study of variance component models in Chapter 9 where the hypotheses being tested will be about variances instead of means.

For data that may not come from a normal population, Conover and Iman (1981) suggest doing analysis of variance on the relative ranks instead of the actual data points. This procedure is also applicable to the other analysis of variance procedures that we will study later. It may be used regardless of the population distribution. However, for data from normal populations, analysis of variance on the observations themselves will probably have higher power than analysis of variance on the relative ranks.

To find the relative ranks, simply order the observations from lowest to highest and assign ranks 1, 2, etc. If there are ties, assign the average of the ranks to each one of them. For example, if we have four observations, one greater than the other and then three ties, we would assign the ranks 1, 2, 3, 4, 6, 6, and 6. The next observation would have rank 8. We now give an example.

**Example 6 Analysis of Variance on Ranks** The data below are taken from those compiled by the National Center for Statistics and Analysis, United States. For five states in different parts of the country, the data represent the number of speeding related fatalities by speed limit in miles per hour during 2003 on non-interstate highways.

| State/Speed Limit | 55 | 50 | 45 |
|---|---|---|---|
| California | 397 | 58 | 142 |
| Florida | 80 | 13 | 150 |
| Illinois | 226 | 3 | 22 |
| New York | 177 | 10 | 23 |
| Washington | 16 | 38 | 15 |

The corresponding table for the relative ranks is

| State/Speed Limit | 55 | 50 | 45 |
|---|---|---|---|
| California | 15 | 9 | 11 |
| Florida | 10 | 3 | 12 |
| Illinois | 14 | 1 | 6 |
| New York | 13 | 2 | 7 |
| Washington | 5 | 8 | 4 |

The analysis of variance table for these relative ranks is

| Source | d.f. | Sum of Squares | Mean Square | $F$ | $p$-value |
|---|---|---|---|---|---|
| Speeds | 2 | 115.6 | 57.8 | 4.22 | 0.041 |
| Error | 12 | 164.4 | 13.7 | | |
| Total | 14 | 280.0 | | | |

At $\alpha = .05$, there is a significant difference in the number of fatalities at different speeds. □

Another non-parametric test based on the ranks is the Kruskal–Wallis Test. For this test of equality of the medians of the distribution functions for each of $k$ random samples, the test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[ R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{73}$$

where for the $i$th sample, $n_i$ is the number of observations and $R_i$ is the sum of the ranks. An equivalent, computationally more convenient form of (73) is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1). \tag{74}$$

The reader may show the equivalence of formulae (73) and (74) in Exercise 15. A reasonable approximation to the distribution of $H$ for sample sizes 5 or larger is the chi-square distribution on $k - 1$ degrees of freedom. Daniel (1990) notes that the power of the Kruskal–Wallis test compares quite favorably with that of the $F$-test for analysis of variance.

**Example 7 Illustration of Kruskal–Wallis Test** For the data of Example 6, we have that $R_1 = 57, R_2 = 23,$ and $R_3 = 40$.

Then using (74),

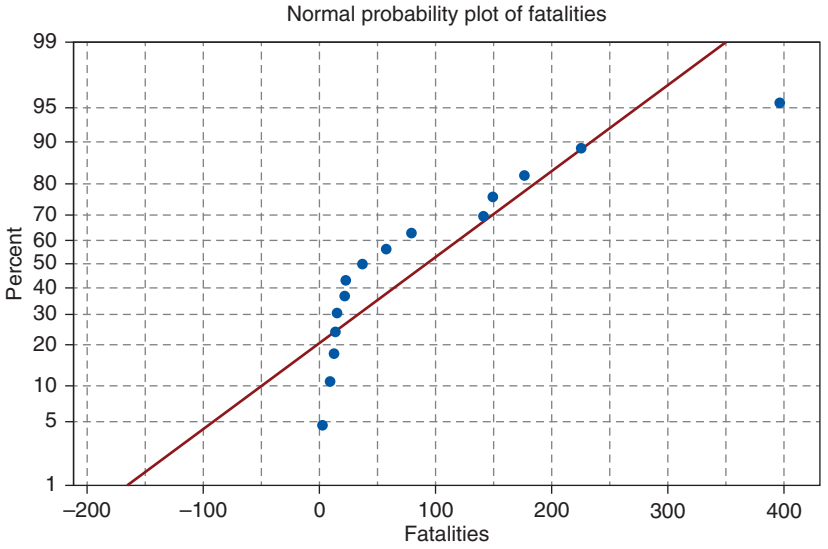
$$H = \frac{12}{(15)(16)} \left[ \frac{57^2}{5} + \frac{23^2}{5} + \frac{40^2}{5} \right] - 3(16) = 5.78.$$

The 95 percentile of the $\chi^2$-distribution is 5.99, so we would fail to reject the hypothesis of equal medians. The $p$-value is .0556. □

Although the test in Example 6 rejects the null hypothesis and the test in Example 7 fails to reject it, the $p$-values in both cases are close to 0.05. By way of comparison for the analysis of variance on the observations, we get the results below.

| Source | d.f. | Sum of Squares | Mean Square | $F$ | $p$-value |
|---|---|---|---|---|---|
| Speeds | 2 | 63194 | 31597 | 3.53 | 0.062 |
| Error | 12 | 107337 | 8945 | | |
| Total | 14 | 170531 | | | |

Again, we would fail to reject the null hypothesis of equal means at the .05 significance level. A big difference in the results of a parametric and non-parametric test might indicate lack of normality. Other tests for normality include making a normal probability plot and determining whether the points are close to a straight line, indicating normality. The normal probability plot below indicates that these data might not be normally distributed.

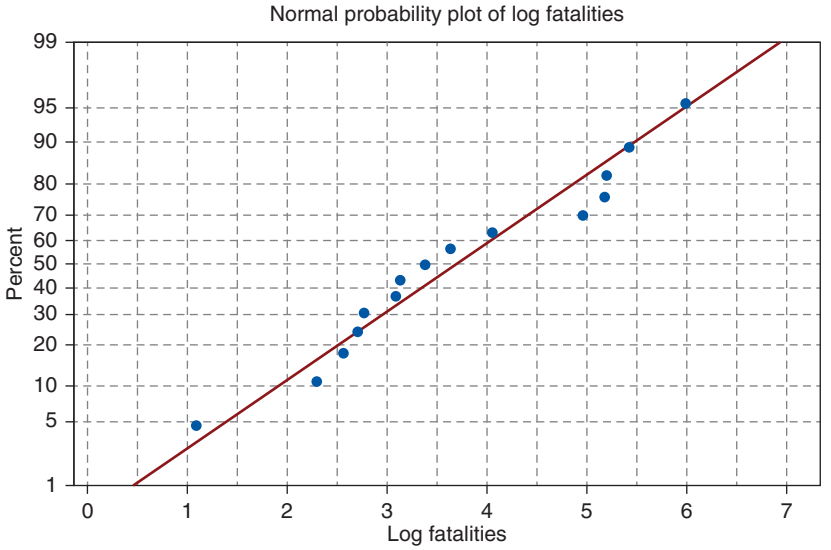



For more on the Kruskal–Wallis test, please see Daniel (1990) and Conover (1998). □

### b. Unequal Variances

Scheffe (1959) observes that "Inequality of variances in the cells of a layout has little effect on the inferences about means but serious effects on inferences about variances of random effects whose kurotosis differs from zero." In particular, for unbalanced

data, it can cause an increase in the type I error and badly invalidate individual confidence intervals. One remedy that sometimes works is to take logarithms of the observations. This may, in certain cases, stabilize the variance. This appears to be the case for the data in Section a.




Bartlett and Kendall (1946), Levene (1960), and Brown and Forsyth (1974a) give tests for the equality of variances. Welch (1951), Brown and Forsyth (1974b), Lee and Ahn (2003), and Rice and Gaines (1989) give modifications of the $F$-test for data with unequal group variances.

We shall give illustrations of Bartlett and Levene's tests for equality of variance.

Then, we will illustrate the modified $F$-test of Brown and Forsyth (1974b) for equality of means where the variances may not be equal, and of Welch (1951).

**(*i*) *Bartlett's Test.*** We are testing the hypothesis that all of the variances are equal versus the alternative that at least one pair of variances is significantly different. The test statistic is

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i-1)\ln s_i^2}{1 + 1/3(k-1)((\sum_{i=1}^{k} 1/(N_i-1)) - 1/(N-k))},$$

where $s_i^2$ is the variance of the $i$th group and $s_p^2$ is the pooled variance given by

$$s_p^2 = \frac{\sum_{i=1}^{k}(N_i-1)s_i^2}{N-k}.$$

The distribution of $T$ is approximately chi-square on $k-1$ degrees of freedom.

**Example 8 Illustration of Bartlett's Test** We use the data from Example 4. We have that

$$s_p^2 = \frac{2(23.16)^2 + 4(56.45)^2 + 3(39.50)^2 + 3(61.10)^2 + 2(29.94)^2 + 3(156.85)^2}{17}$$
$$= 6194.01$$

and

$$T = \frac{17\ln(6194.01) - 4\ln(23.16) - 8\ln(56.45) - 6\ln(39.5) - 6\ln(61.1) - 4\ln(29.94) - 6\ln(156.85)}{1 + (1/15)(1/2 + 1/4 + 1/3 + 1/3 + 1/2 + 1/3 - 1/17)}$$
$$= \frac{12.9348}{1.14608} = 11.286 > 11.1$$

We reject the hypothesis of equal variances at $\alpha = .05$. The $p$-value is .046. □

Unfortunately, Bartlett's test is very heavily dependent on the data being normally distributed.

***(ii) Levene's Test.*** This consists of doing an analysis of variance on the absolute values of the differences between the observations and either the mean, the median, or the trimmed mean. It is less dependent on whether the data are from a normal population than Bartlett's test. However, if the data come from a normal population Bartlett's test is more powerful. Thus, it is the preferred test when there is strong evidence of a normal population. We shall demonstrate Levene's test for the data of Example 4 using the medians. The original Levene's test used means. The modification for medians and trimmed means is due to Brown and Forsyth (1974a) and is more robust.

**Example 9 Illustration of Levene Test Using Data from Example 4** The transformed data that we need to do the analysis of variance of are below.

| State/Speed Limit | 55 | 50 | 45 | 40 | 35 | < 35 |
|---|---|---|---|---|---|---|
| California | 195.5 | 47 | 119.5 | 68.5 | 104 | – |
| Florida | – | – | – | – | 11 | 0 |
| Illinois | 24.5 | 7 | 0.5 | 8.5 | 0 | 13 |
| New York | 24.5 | 0 | 0.5 | 8.5 | 46 | – |
| Washington | 185.5 | – | 7.5 | 20.5 | 16 | 32 |

The resulting analysis of variance table is given below.

| Source | d.f. | Sum of Squares | Mean Square | $F$ | $p$ |
|---|---|---|---|---|---|
| Speeds | 5 | 22625 | 4525 | 1.57 | 0.223 |
| Error | 17 | 49142 | 2892 | | |
| Total | 22 | 71768 | | | |

In this instance, we fail to reject the hypothesis of equal variances. When a parametric test that assumes that the data come from a normal population gives different results from a non-parametric test, there is suspicion that the data are not normal. □

**(*iii*) *Welch's (1951) F-test.*** The statistic used for the Welch $F$-test is

$$F_w = \frac{MS_M^W}{1 + 2\mathbf{\Lambda}(a-2)/3}, \tag{75}$$

where

$$MS_M^w = \frac{\sum_{i=1}^a w_i(\bar{y}_{i.} - \bar{y}_{..}^w)^2}{a-1}$$

with $w_i = n_i/s_i^2$, $\bar{y}_{..}^w = \sum_{i=1}^a w_i \bar{y}_{i.} / \sum_{i=1}^a w_i$, and

$$\mathbf{\Lambda} = \frac{3\sum_{i=1}^a \left(1 - \frac{w_i}{\sum_{i=1}^a w_i}\right)^2 / (n_i - 1)}{a^2 - 1}.$$

The model degrees of freedom are $a-1$ and the residual degrees of freedom are $1/\mathbf{\Lambda}$.

**Example 10 Welch *F*-ratio for Data of Example 4** We have that

$$w_1 = \frac{3}{(23.16)^2} = 0.00559299, \quad w_2 = \frac{5}{(56.45)^2} = 0.00156907, \quad w_3 = \frac{4}{(39.50)^2} = 0.00256369,$$

$$w_4 = \frac{4}{(61.10)^2} = 0.00107146, \quad w_5 = \frac{3}{(29.94)^2} = 0.00334671, \quad w_6 = \frac{4}{(156.85)^2} = 0.00162589$$

and

$$w = 0.00559299 + 0.00156907 + 0.00256369 + 0.00107146 + 0.00334671 + 0.00162589 = 0.057698$$

Then, by substitution in $\bar{y}_{..}^{w} = \sum_{i=1}^{a} w_i \bar{y}_{i.} / \sum_{i=1}^{a} w_i$, we have that $\bar{y}_{..}^{w} = 30.1524$ and $MS_M^w = \frac{\sum_{i=1}^{a} w_i (\bar{y}_{i.} - \bar{y}_{..}^{w})^2}{a-1} = 36.2406$. The degrees of freedom, d.f. $= 1/\Lambda = 1/0.129663 = 7.71231$. Finally, from (75), our $F$-statistic is

$$F = 26.9263.$$

Comparing this to $F_{.05,5,7} = 3.971$, we reject the hypothesis of equal variances. □

**(*iv*) *Brown–Forsyth (1974b) Test.*** The test statistic is given by

$$F = \frac{\sum_{i=1}^{a} n_i (\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{a} (1 - n_i/N) s_i^2}, \tag{76}$$

where $N = \sum_{i=1}^{a} n_i$ and $s_i^2 = \sum_{j=1}^{n_i} (\overleftarrow{y}_{ij} - \bar{y}_{i.})^2/(n_i - 1)$. This approximate $F$-distribution has degrees of freedom given by

$$\text{d.f.} = \frac{1}{\sum_{i=1}^{a} c_i^2/(n_i - 1)},$$

with

$$c_i = \frac{(1 - n_i/N) s_i^2}{\sum_{i=1}^{a} (1 - n_i/N) s_i^2}$$

The numerator of (76) is the usual between sum of squares for one-way analysis of variance.

**Example 11 Brown–Forsyth (1974b) *F*-test for Data of Example 4** From the SAS output, the sum of squares for the numerator of (76) is 78,240.9449.

The denominator is $(\frac{1}{23})(20(23.1589)^2 + 18(56.4517)^2 + 19(39.5055)^2 + 19(61.1037)^2 + 20(29.9388)^2 + 19(156.8502)^2) = 28436.8$.

The resulting $F$-statistic is

$$F = \frac{78240.9499}{28436.8} = 2.7514.$$

We need to approximate the degrees of freedom. We have,

$$c_1 = \frac{20(23.1589)^2}{654046} = 0.0164005, \quad c_2 = \frac{18(56.4517)^2}{654046} = 0.0877038,$$
$$c_3 = \frac{19(39.5011)^2}{654046} = 0.0453277, \quad c_4 = \frac{19(61.1037)^2}{654046} = 0.108463,$$
$$c_5 = \frac{20(29.9388)}{654046} = 0.0274088, \quad c_6 = \frac{19(156.8502)}{654046} = 0.714696.$$

Then,

$$\text{d.f.} = \frac{1}{c_1^2/2 + c_2^2/4 + c_3^2/3 + c_4^2/3 + c_5^2/2 + c_6^2/3} = 5.64 \simeq 6.$$

We would fail to reject the hypothesis of equal means at $\alpha = .05$ because $F_{.05,5,6} = 4.4$. The approximate $p$-value is 0.125.

### c. Non-independent Observations

For the three basic assumptions for analysis of variance, violation of the independence observation is the most serious. To illustrate this, following the discussion in Scheffe (1959), we consider $n$ observations $y_i$ from a normal population that are serially correlated. Consideration of this special case will illustrate the issues involved while simplifying the discussion. Assume that $E(y_i) = \mu$ and $var(y_i) = \sigma^2$, the correlation coefficient if $y_i$ and $y_{i+1}$ is $\rho$ for $i = 1, \ldots, n-1$ and that all other correlation coefficients are zero. Then, we have that

$$E(\bar{y}) = \mu, \tag{77a}$$

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n}\left[1 + 2\rho\left(1 - \frac{1}{n}\right)\right] \tag{77b}$$

and

$$E(s^2) = \sigma^2\left(1 - \frac{2\rho}{n}\right) \tag{77c}$$

See Exercise 11.

If in (74b), we drop the term $1/n^2$, we observe that the random variable

$$t = \frac{(\bar{y} - \mu)}{s/\sqrt{n}}$$

which follows a Student $t$-distribution for small $n$ is asymptotically $N(0, 1 + \rho)$. Then, the probability of a large sample confidence interval with confidence coefficient $1 - \alpha$ not covering the true mean $\mu$ is given by the integral

$$\text{I} = \frac{2}{\sqrt{2\pi}} \int_{z_{\alpha/2}/(1+2\rho)}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

As $\rho \to -\frac{1}{2}, I \to 0$. As $\rho \to \frac{1}{2}, I \to 0.17$ for $\alpha = .05$. This illustrates how the effect of serial correlation on inferences about means can be quite serious.

Schauder and Schmid (1986) investigate one-way analysis of variance assuming that within each group the correlation between any two observations is the same. They observe that it is highly non-robust with respect to positive within group correlation. Positive-within-group correlation strongly increases the level of significance of the test. Negative-within-group correlation renders the test conservative.

Adke (1986) observes that for the most part, analysis of variance is invalid when their observations within groups are correlated. Independence is needed.

To determine whether the observations are independent or not, you can make a plot of the residuals in the order in which the observations are collected and see if there are any patterns. The Durbin–Watson test can also be performed.

The test statistic for the Durbin–Watson test is for residuals $r_i$

$$d = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2} \tag{78}$$

assuming that the observations are in the time order in which they occurred. The values are tabulated according to the number of regressors as lower and upper values $d_L$ and $d_u$ for different levels of significance. When the computed value of the Durbin–Watson statistic is less than $d_L$, the hypothesis of independence may be rejected and there may be a serial correlation between successive observations. If on the other hand, the computed Durbin–Watson statistic is greater than $d_u$, there is insufficient evidence of non-independence. If the statistic falls between $d_u$ and $d_L$, the test is inconclusive. Tables are readily available online. One website is http://www.stat.ufl.edu/~winner/tables/DW_05.pdf.

The $p$-values of the Durbin–Watson test may be obtained using SAS and R.

## 6. THE TWO-WAY NESTED CLASSIFICATION

This section will consider the two-way nested classification. We shall give the form of the linear model, the normal equations and their solutions, the analysis of variance, the estimable functions, and show how to formulate tests of hypotheses. As a case in point, we will use an example from Chapter 4 that describes a student opinion poll of instructions use of a computing facility in courses in English, Geology, and Chemistry. Table 6.5 contains partial data from such a poll. The data are from a two-way nested classification. We now describe its analysis in the subsections that follow.

**TABLE 6.5 Student Opinion Poll of Instructor's Classroom Use of Computer Facility**

| | | Observations | | | |
|---|---|---|---|---|---|
| Course | Section of Course | Individual | Total | Number | Mean |
| English | 1 | 5 | 5 | (1) | 5 |
| | 2 | 8, 10, 9 | 27 | (3) | 9 |
| | | | Total 32 | (4) | 8 |
| Geology | 1 | 8, 10 | 18 | (2) | 9 |
| | 2 | 6, 2, 1, 3 | 12 | (4) | 3 |
| | 3 | 3, 7 | 10 | (2) | 5 |
| | | | | (8) | 5 |
| | | | | (12) | 6 |

### a. Model

As suggested in Chapter 4, a suitable model is

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}. \tag{79}$$

The $y_{ijk}$ is the $k$th observation in the $j$th section of the $i$th course. The term $\mu$ is a general mean. The effect due to the $i$th course is $\alpha_i$. The $\beta_{ij}$ is the effect due to $j$th section of the $i$th course. The usual error term is $e_{ijk}$.

There are $a$ levels of the $\alpha$-factor (courses). For the data of Table 6.5, $i = 1, 2, \ldots, a$ with $a = 2$. For $b_i$ levels for the $\beta$-factor nested within the $\alpha$-factor (sections nested within courses), $j = 1, 2, \ldots, b_i$, with $b_1 = 2$ and $b_2 = 3$ in the example. Furthermore, for $n_{ij}$ observations in the $j$th section of the $i$th course, $k = 1, 2, \ldots, n_{ij}$. The values of the $n_{ij}$ in Table 6.5 are those in the penultimate column thereof. This column also shows the values of $n_{i.} = \sum_{j=1}^{b_i} n_{ij}$ and $n.. = \sum_{i=1}^{a} n_{i.}$ We have that $n_{11} = 1, n_{12} = 3, n_{1.} = 4$, and $n.. = 12$. The table also contains the corresponding totals and means of $y_{ijk}$.

### b. Normal Equations

For the 12 observations of Table 6.6, the equations of the model (79) are

$$\begin{bmatrix} 5 \\ 8 \\ 10 \\ 9 \\ 8 \\ 10 \\ 6 \\ 2 \\ 1 \\ 3 \\ 3 \\ 7 \end{bmatrix} = \begin{bmatrix} y_{111} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{224} \\ y_{231} \\ y_{232} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{121} \\ e_{122} \\ e_{123} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{223} \\ e_{224} \\ e_{231} \\ y_{232} \end{bmatrix} \tag{80}$$

**TABLE 6.6 Analysis of Variance for the Data of Table 5.5**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean | $1 = 1$ | $R(\mu) = 432$ | 432 | $F(\mathrm{M}) = 1163$ |
| Model, after mean | $b - 1 = 4$ | $R(\alpha,\beta{:}\alpha\|\mu) = 84$ | 21 | $F(R_{\mu}) = 5.65$ |
| Residual | $N - b = 7$ | SSE = 26 | 3.714 | |
| Total | | SST = 542 | | |

Writing $\mathbf{X}$ for the $12 \times 8$ matrix of 0's and 1's, it follows that the normal equations are $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ are

$$\left[\begin{array}{c:cc:cc:cc} 12 & 4 & 8 & 1 & 3 & 2 & 4 & 2 \\ \hdashline 4 & 4 & 0 & 1 & 3 & 0 & 0 & 0 \\ 8 & 0 & 8 & 0 & 0 & 2 & 4 & 2 \\ \hdashline 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 3 & 0 & 0 & 0 \\ \hdashline 2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 & 4 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 0 & 2 \end{array}\right] \left[\begin{array}{c} \mu^\circ \\ \hdashline \alpha_1^\circ \\ \alpha_2^\circ \\ \hdashline \beta_{11}^\circ \\ \beta_{12}^\circ \\ \hdashline \beta_{21}^\circ \\ \beta_{22}^\circ \\ \beta_{23}^\circ \end{array}\right] = \left[\begin{array}{c} 72 \\ \hdashline 32 \\ 40 \\ \hdashline 5 \\ 27 \\ \hdashline 18 \\ 12 \\ 10 \end{array}\right] = \left[\begin{array}{c} y_{...} \\ \hdashline y_{1..} \\ y_{2..} \\ \hdashline y_{11.} \\ y_{12.} \\ \hdashline y_{21.} \\ y_{22.} \\ y_{23.} \end{array}\right] \tag{81}$$

The general form of these equations is

$$\left[\begin{array}{c:cc:cc:cc} n_{..} & n_{1.} & n_{2.} & n_{11} & n_{12} & n_{21} & n_{22} & n_{23} \\ \hdashline n_{1.} & n_{1.} & 0 & n_{11} & n_{12} & 0 & 0 & 0 \\ n_{2.} & 0 & n_{2.} & 0 & 0 & n_{21} & n_{22} & n_{23} \\ \hdashline n_{11} & n_{11} & 0 & n_{11} & 0 & 0 & 0 & 0 \\ n_{12} & n_{12} & 0 & 0 & n_{12} & 0 & 0 & 0 \\ \hdashline n_{21} & 0 & n_{21} & 0 & 0 & n_{21} & 0 & 0 \\ n_{22} & 0 & n_{22} & 0 & 0 & 0 & n_{22} & 0 \\ n_{23} & 0 & n_{23} & 0 & 0 & 0 & 0 & n_{23} \end{array}\right] \left[\begin{array}{c} \mu^\circ \\ \hdashline \alpha_1^\circ \\ \alpha_2^\circ \\ \hdashline \beta_{11}^\circ \\ \beta_{12}^\circ \\ \hdashline \beta_{21}^\circ \\ \beta_{22}^\circ \\ \beta_{23}^\circ \end{array}\right] = \left[\begin{array}{c} y_{...} \\ \hdashline y_{1..} \\ y_{2..} \\ \hdashline y_{11.} \\ y_{12.} \\ \hdashline y_{21.} \\ y_{22.} \\ y_{23.} \end{array}\right] \tag{82}$$

The partitioning shown in (82) suggests how more levels of factors would be incorporated into the normal equations.

### c. Solving the Normal Equations

The square matrix $\mathbf{X}'\mathbf{X}$ in equations (81) and (82) has order 8 and rank 5. To see that its rank is 5, observe that rows 2 and 3 sum to row 1, rows 4 and 5 sum to row 2, and rows 6, 7, and 8 sum to row 3. Hence $r(\mathbf{X}'\mathbf{X}) = 8 - 3 = 5$. For the general two-way nested classification, $\mathbf{X}'\mathbf{X}$ has rank $b$, the number of subclasses. This holds true because its order $p$ is, for $a$ levels of the main classification (courses in our example), $p = 1 + a + b$.

However, the rows corresponding to the $\alpha$-equations add to that of the $\mu$-equations (1 dependency) and the rows corresponding to the $\beta$-equations in each $\alpha$-level add to the row for that $\alpha$-equation ($a$ dependencies, linearly independent of the first one). Therefore, $r = r(\mathbf{X}'\mathbf{X}) = 1 + a + b. - (1 + a) = b$. Hence by (4), the normal equations can be solved by putting $p - r = 1 + a$ elements of $\mathbf{b}^\circ$ equal to zero. From the nature of (81) and (82), it appears that the easiest $1 + a$ elements of $\mathbf{b}^\circ$ to set equal to zero are $\mu^\circ$ and $\alpha_1^\circ, \alpha_2^\circ, \ldots, \alpha_a^\circ$. As a result, the other elements of $\mathbf{b}^\circ$ are

$$\beta_{ij}^\circ = \bar{y}_{ij.} \quad \text{for all } i \text{ and } j. \tag{83}$$

Thus, a solution to the normal equations is

$$\mathbf{b}^{\circ\prime} = \left[\mathbf{0}'_{1\times(1+a)} \quad \bar{\mathbf{y}}'\right], \tag{84}$$

where $\bar{\mathbf{y}}'$ is the row vector of cell means. Note that the solution in (84) is not unique.

For the case of the example, we see from Table 6.5 that

$$\bar{\mathbf{y}}' = \begin{bmatrix} 5 & 9 & 9 & 3 & 5 \end{bmatrix} \tag{85a}$$

and the corresponding solution to the normal equation is

$$\mathbf{b}^{\circ\prime} = \begin{bmatrix} 0 & 0 & 0 & 5 & 9 & 9 & 3 & 5 \end{bmatrix}. \tag{85b}$$

The corresponding generalized inverse of $\mathbf{X}'\mathbf{X}$ is

$$\mathbf{G} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(1/n_{ij}) \end{bmatrix} \quad \text{for} \quad i = 1, \ldots, a, j = 1, 2, \ldots, b_i, \tag{86}$$

where $\mathbf{D}(1/n_{ij})$ for the example is diagonal, with non-zero elements $1, \frac{1}{3}, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{2}$.

### d. Analysis of Variance

The sums of squares for the analysis of variance for this model for the example of Table 6.5 are as follows:

$$\begin{aligned} R(\mu) = \text{SSM} &= n_{\cdot\cdot} \bar{y}^2_{\cdots} = 12(6)^2 = 432; \\ R(\mu, \alpha, \beta : \alpha) = \text{SSR} &= \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{a} \sum_{j=1}^{b_i} \frac{y^2_{ij\cdot}}{n_{ij}} \\ &= \frac{5^2}{1} + \frac{27^2}{3} + \frac{18^2}{2} + \frac{12^2}{4} + \frac{10^2}{2} = 516; \\ R(\alpha, \beta : \alpha) &= R(\mu, \alpha, \beta : \alpha) - R(\mu) = 516 - 432 = 84; \\ \text{SST} &= 5^2 + 8^2 + \cdots + 3^2 + 7^2 = 542; \\ \text{SSE} &= \text{SST} - R(\mu, \alpha, \beta : \alpha) = 542 - 516 = 26. \end{aligned}$$

Hence the analysis of variance table, in the style of Table 6.4 is that shown in Table 6.6. Since $F(M) = 116.3$, we reject the hypothesis $H$: $E(\bar{y}) = 0$ at $\alpha = .05$ because $F(M) > F_{.05,1,7} = 5.59$. Likewise, we also reject the hypothesis at $\alpha = .05$ that the model $E(y_{ijk}) = \mu + \alpha_i + \beta_{ij}$ of (79) does not account for more variation in the $y$ variable than does the model $E(y_{ijk}) = \mu$ because $F(R_{\text{m}}) > F_{.05,4,7} = 4.12$.

Suppose that we fit the one-way classification model

$$y_{ijk} = \mu + \alpha_i + e_{ijk}$$

to the data of Table 6.5. Then, as in (37) and (76), the reduction for fitting this model is

$$R(\mu, \alpha) = \sum_{i=1}^{a} \frac{y_{i..}}{n_{i.}} = \frac{32^2}{4} + \frac{40^2}{8} = 456.$$

Hence,

$$R(\beta : \alpha | \mu, \alpha) = R(\mu, \alpha, \beta : \alpha) - R(\mu, \alpha) = 516 - 546 = 60,$$

and

$$R(\alpha | \mu) = R(\mu, \alpha) - R(\mu) = 456 - 432 = 24.$$

As a result, we can divide $R(\alpha, \beta : \alpha | \mu)$ of Table 6.6 into two portions. Observe that

$$\begin{aligned} 84 &= R(\alpha, \beta : \alpha | \mu) = R(\mu, \alpha, \beta : \alpha) - R(\mu) \\ &= R(\mu, \alpha, \beta : \alpha) - R(\mu, \alpha) + R(\mu, \alpha) - R(\mu) \\ &= R(\alpha, \beta : \alpha | \mu) + R(\alpha | \mu) \\ &= 60 + 24. \end{aligned}$$

We see the result of doing this in Table 6.7 where the $F$-statistic is

$$F(\alpha | \mu) = \frac{R(\alpha, \mu)}{(a-1)MSE} = 6.46 > F_{.05,1,7} = 5.59. \tag{87}$$

This tests the significance of fitting $\alpha$ after $\mu$. Furthermore,

$$F(\beta : \alpha | \mu, \alpha) = \frac{R(\beta : \alpha | \mu, \alpha)}{(b. - a)\text{MSE}} = 5.39 > F_{.05,3,7} = 4.35 \tag{88}$$

tests the significance of fitting $\beta : \alpha$ after $\alpha$ and $\mu$. From (87) and (88), we conclude that fitting $\alpha$ after $\mu$ as well as $\beta : \alpha$ after $\alpha$ and $\mu$ accounts for the variation in the $y$ variable.

**TABLE 6.7 Analysis of Variance for Data of Table 6.5 (Two-Way Nested Classification)**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean, $\mu$ | $1 = 1$ | $R(\mu) = 432$ | 432 | 116.3 |
| $\alpha$ after $\mu$ | $\alpha - 1 = 1$ | $R(\alpha \mid \mu) = 24$ | 24 | 6.46 |
| $\beta$:$\alpha$ after $\mu$ and $\alpha$ | $b - \alpha = 3$ | $R(\beta$:$\alpha \mid \mu, \alpha) = 60$ | 20 | 5.39 |
| Residual | $N - b = 7$ | SSE = 26 | 3.714 | |
| Total | $N = 12$ | SST = 542 | | |

### e. Estimable Functions

Applying the general theory of estimability to any design models involves many of the points detailed in Section 2e for the one-way classification. We will not repeat these details in what follows.

The expected value of any observation is estimable. Thus, from (83) and (85), $\mu + \alpha_i + \beta_{ij}$ is estimable with b.l.u.e. $\mu + \alpha_i^\circ + \beta_{ij}^\circ = \bar{y}_{ij.}$. Table 6.8 contains this result and linear combinations thereof. An illustration of one of them is, using (85),

$$\widehat{\beta_{11} - \beta_{12}} = 5 - 9 = -4.$$

Its variance is

$$v(\widehat{\beta_{11} - \beta_{12}}) = \sigma^2 \left( \frac{1}{1} + \frac{1}{3} \right) = \frac{4\sigma^2}{3}.$$

From Table 6.6, an unbiased estimate of this variance is $4\hat{\sigma}^2/3 = 4(\text{MSE})/3 = 4(3.714)/3 = 4.952$. Typically, one uses the values $1/b_i$ or $n_{ij}/n_{i.}$ for $w_{ij}$ in the last two rows of Table 6.8. For example, using $1/b_i$ using (85) again, we have, for example, that the b.l.u.e. of

$$\alpha_1 - \alpha_2 + \frac{1}{2}(\beta_{11} + \beta_{12}) - \frac{1}{3}(\beta_{21} + \beta_{22} + \beta_{23}) \tag{89}$$

for the data in Table 6.5 has the estimate of

$$\frac{1}{2}(5 + 9) - \frac{1}{3}(9 + 3 + 5) = \frac{4}{3}.$$

An estimate of the variance of the b.l.u.e. is

$$\sigma^2 \left[ \left( \frac{1}{2} \right)^2 \left( \frac{1}{1} + \frac{1}{3} \right) + \left( \frac{1}{3} \right)^2 \left( \frac{1}{2} + \frac{1}{4} + \frac{1}{2} \right) \right] = \frac{17}{36}\sigma^2.$$

Note that in Table 6.8, none of the linear functions $\mu$, $\mu + \alpha_i$, and $\alpha_i$ are estimable.

**TABLE 6.8 Estimable Functions in the Two-Way Nested Classification $y_{ij} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$**

| Estimable Function | b.l.u.e. | Variance of b.l.u.e. |
|---|---|---|
| $\mu + \alpha_i + \beta_{ij}$ | $\bar{y}_{ij.}$ | $\frac{\sigma^2}{n_{ij}}$ |
| $\beta_{ij} - \beta_{ij'}$ for $j \neq j'$ | $\bar{y}_{ij.} - \bar{y}_{ij'.}$ | $\sigma^2 \left( \frac{1}{n_{ij}} + \frac{1}{n_{ij'}} \right)$ |
| $\mu + \alpha_i + \sum_{j=1}^{b_i} w_{ij}\beta_{ij}$ | $\sum_{j=1}^{b_i} w_{ij}\bar{y}_{ij}$ | $\sigma^2 \left( \sum_{j=1}^{b_i} \frac{w_{ij}^2}{n_{ij}} \right)$ |
| for $\sum_{j=1}^{b_i} w_{ij} = 1$ | | |
| $\alpha_i - \alpha_{i'} + \sum_{j=1}^{b_i} w_{ij}\beta_{ij} - \sum_{j=1}^{b_{i'}} w_{i'j}\beta_{i'j}$ | | |
| for $\sum_{j=1}^{b_i} w_{ij} = 1 = \sum_{j=1}^{b_{i'}} w_{i'j}$ | $\sum_{j=1}^{b_i} w_{ij}\bar{y}_{ij} - \sum_{j=1}^{b_i} w_{i'j}\bar{y}_{i'j}$ | $\sigma^2 \left( \sum_{j=1}^{b_i} \frac{w_{ij}^2}{n_{ij}} + \sum_{j=1}^{b_{i'}} \frac{w_{i'j}^2}{n_{i'j}} \right)$ |

### f. Tests of Hypothesis

The estimable functions of Table 6.8 form the basis of testable hypotheses. The $F$-statistic for testing the null hypothesis that any one of the functions in Table 6.8 is zero is the square of its b.l.u.e. divided by that b.l.u.e.'s variance with $\hat{\sigma}^2$ replacing $\sigma^2$. Under the null hypothesis, such a statistic has the $F_{1,N-b}$-distribution. Its square root has the $t_{N-b}$-distribution. Thus, we can use the statistic

$$F = \frac{(\bar{y}_{ij.} - \bar{y}_{ij'.})^2}{\hat{\sigma}^2(1/n_{ij} + 1/n_{ij'})}$$

or equivalently $\sqrt{F}$ to test the hypothesis that $\beta_{ij} = \beta_{ij'}$.

The hypothesis $H$: $\beta_{i1} = \beta_{i2} = \cdots = \beta_{ib_i}$ is of special interest. It is the hypothesis of equal $\beta$'s within each $\alpha$ level. By writing it in the form $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$, it can be shown that the resulting $F$-statistic of (21) is $F(\beta : \alpha | \mu, \alpha)$ that was given in (88) and used in Table 6.8. Recall that in (88), $F(\beta : \alpha | \mu, \alpha)$ was given as the statistic for testing the significance of $\beta : \alpha$ after $\mu$ and $\alpha$. Equivalently, this statistic can be used to test the hypothesis of equalities of the $\beta$'s within each $\alpha$ level.

**Example 12 Test of the Equalities of the $\beta$'s Within Each $\alpha$ Level for the Data of Table 6.5** Carrying out this test for the data of Table 6.5 involves using

$$\mathbf{K}' = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}. \tag{90}$$

Using $\mathbf{b}^{\circ\prime}$ of (85) and $\mathbf{G}$ implicit in (86) gives

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} -4 \\ 6 \\ 4 \end{bmatrix} \text{ and } (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \begin{bmatrix} \frac{4}{3} & 0 & 0 \\ 0 & \frac{3}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{3}{4} & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & \frac{3}{2} \end{bmatrix}.$$

Then $Q$ of (21) is

$$Q = \begin{bmatrix} -4 & 6 & 4 \end{bmatrix} \begin{bmatrix} \frac{3}{4} & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} -4 \\ 6 \\ 4 \end{bmatrix} = 60 = R(\beta : \alpha, | \mu, \alpha)$$

of Table 6.7. Thus the $F$-value is $60/3\hat{\sigma}^2 = 60/3(3.714) = 5.39$. □

**Example 13 Another Hypothesis Test for an Estimable Function of the Type in the Last Line of Table 6.8** Consider the hypothesis

$$H\colon \mathbf{k}'\mathbf{b} = 0 \quad \text{for} \quad \mathbf{k}' = \left[0 \quad 1 \quad -1 \quad \tfrac{1}{4} \quad \tfrac{3}{4} \quad -\tfrac{2}{8} \quad -\tfrac{4}{8} \quad -\tfrac{2}{8}\right]. \tag{91}$$

We have that

$$\mathbf{k}'\mathbf{b} = \alpha_1 - \alpha_2 + \frac{1}{4}(\beta_{11} + 3\beta_{12}) - \frac{1}{8}(2\beta_{21} + 4\beta_{22} + 2\beta_{23}).$$

This is an estimable function of the type found in the last line of Table 6.8 with $w_{ij} = n_{ij}/n_{i.}$. From (91), (85), and (86), we have that,

$$\mathbf{k}'\mathbf{b} = 3 \quad \text{and} \quad \mathbf{k}'\mathbf{G}\mathbf{k} = \frac{3}{8}.$$

Thus, by (21) the numerator sum of squares for testing the hypothesis in (91) is

$$Q = 3^2\left(\frac{8}{3}\right) = 24 = R(\alpha|\mu) \text{ of Table 6.7.} \tag{92}$$

□

The result obtained in (92) is no accident. Although $R(\alpha|\mu)$ is as indicated in (87), the numerator sum of squares for testing the fit of $\alpha$ after $\mu$, it is also the numerator sum of squares for testing

$$H\colon \alpha_i + \sum_{j=1}^{b_i} \frac{n_{ij}\beta_{ij}}{n_{i.}} = \alpha_{i'} + \sum_{j=1}^{b_i} \frac{n_{i'j}\beta_{i'j}}{n_{i'.}} \quad \text{for all} \quad i \neq i'. \tag{93}$$

Furthermore, in the sense of (62), the hypothesis in (93) is orthogonal to

$$H\colon \beta_{ij} = \beta_{ij'} \quad \text{for} \quad j \neq j', \text{ within each } i. \tag{94}$$

The hypothesis $H\colon \mathbf{k}'\mathbf{b} = 0$ that uses $\mathbf{k}'$ in (91) and the hypothesis $H\colon \mathbf{K}'\mathbf{b} = 0$ that uses $\mathbf{K}'$ of (90) are examples of (93) and (94). Every row of $\mathbf{k}'$ and every row of $\mathbf{K}'$ satisfy (62) ($\mathbf{k}'\mathbf{G}\mathbf{K} = 0$). Furthermore, when we test (93) by using (21), we will find that $F(H)$ will reduce to $F(\alpha|\mu)$ as exemplified in (92). Hence, $F(\alpha|\mu)$ tests (93) with numerator sum of squares $R(\alpha|\mu)$. Likewise, $F(\beta : \alpha|\mu, \alpha)$ tests (94) with numerator sum of squares $R(\beta : \alpha|\mu, \alpha)$. The two numerator sums of squares $R(\alpha|\mu)$ and $R(\beta : \alpha|\mu, \alpha)$ are statistically independent. This can be established by expressing each sum of squares as quadratic forms in $y$ and applying Theorem 7 of Chapter 2 (see Exercise 13).

We can also appreciate the equivalence of the $F$-statistic for testing (93) and $F(\alpha|\mu)$ by noting in (93) if $\beta_{ij}$ did not exist then (93) would represent

$$H\text{: all } \alpha'\text{s equal (in the absence of } \beta'\text{s).}$$

This, indeed, is the context of earlier interpreting $F(\alpha|\mu)$ as testing $\alpha$ after $\mu$.

### g. Models that Include Restrictions

The general effect of having restrictions as part of the model has been discussed in Section 6 of Chapter 5 and illustrated in detail in Section 2h of the present chapter. The points made there apply equally as well here. Restrictions that involve non-estimable functions of the parameters affect the form of functions that are estimable and hypotheses that are testable. The restrictions of particular interest here are $\sum_{j=1}^{b_i} w_{ij}\beta_{ij} = 0$ with $\sum_{j=1}^{b_i} w_{ij} = 1$ for all $i$ because then we see from Table 6.8 that $\mu + \alpha_i$ and $\alpha_i - \alpha_{i'}$ are estimable and hypotheses about them are testable. Suppose that the $w_{ij}$ of the restrictions are $n_{ij}/n$, so that the restrictions are $\sum_{j=1}^{b} n_{ij}\beta_{ij} = 0$ for all $i$. For this case, (93) becomes $H$: all $\alpha_i'$s equal. Then, as we have just shown (93) is tested by $F(\alpha|\mu)$. This $F$-statistic is independent of $F(\beta : \alpha|\mu, \alpha)$ that tests $H$: all $\beta'$s equal within each $\alpha$ level. On the other hand, suppose that the $w_{ij}$ of the restrictions are not $n_{ij}/n$ but instead take some other form where $\sum_{j=1}^{b_i} w_{ij} = 1$ for all $i$.

For example, suppose we have $w_{ij} = 1/b_i$. For this case, we can still test the hypothesis $H$: all $\alpha_i'$s equal. However, the $F$-statistic will not be equal to $F(\alpha|\mu)$, nor will its numerator be independent of $F(\beta : \alpha|\mu, \alpha)$.

### h. Balanced Data

The position with balanced data ($n_{ij} = n$ for all $i$ and $j$ and $b_i = b$ for all $i$) is akin to that of the one-way classification discussed in Section 2i earlier. Applying "constraints" $\sum_{j=1}^{b_i} \beta_{ij}^{\circ} = 0$ for all $i$ and $\sum_{i=1}^{a} \alpha_i^{\circ} = 0$ to the normal equations lead to easy solutions thereof: $\mu^{\circ} = \bar{y}_{...}$, $\alpha_i^{\circ} = \bar{y}_{i..} - \bar{y}_{...}$, and $\beta_{ij}^{\circ} = \bar{y}_{ij.} - \bar{y}_{...}$, as is found in many texts. Other results are unaltered. For example, the estimable functions and their b.l.u.e.'s are the same.

When restrictions paralleling the constraints are taken as part of the model, $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b_i} \beta_{ij} = 0$ for all $i$, the effect is to make $\mu, \alpha_i$, and $\beta_{ij}$ individually estimable with b.l.u.e.'s $\hat{\mu} = \bar{y}_{...}$, $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$, and $\hat{\beta}_{ij} = \bar{y}_{ij.} - \bar{y}_{...}$. As was the case with the one-way classification, rationalization of such restrictions is opportune. The $\alpha_i$'s are defined as deviations from their mean. Likewise, the $\beta_{ij}$'s are the deviations from their within $\alpha$-level means.

## 7. NORMAL EQUATIONS FOR DESIGN MODELS

Models of the type described here, in Chapter 4 and later on in this book are sometimes called *design* models (see, for example, Graybill (1976)). We will now characterize some general properties of the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}$ of design models using (81) as a case in point.

The following are some general properties of normal equations.

1. There is one equation corresponding to each effect of a model.
2. The right-hand side of any equation (the element of $\mathbf{X}'\mathbf{y}$) is the sum of all observations that contain in their model, a specific effect. For example, the right-hand side of the first equation in (81) is the sum of all observations that contain $\mu$.
3. The left-hand side of each equation is the expected value of the right-hand side with $\mathbf{b}$ replaced by $\mathbf{b}^\circ$.

As a result of the above observations, the first equation in (82) corresponds to $\mu$.

Its right-hand side is $y_{...}$. Its left-hand side is $E(y_{...})$ with $\mathbf{b}$ therein replaced by $\mathbf{b}^\circ$. Hence, the equation is as implied in (81),

$$12\mu^\circ + 4\alpha_1^\circ + 8\alpha_2^\circ + \beta_{11}^\circ + 3\beta_{12}^\circ + 2\beta_{21}^\circ + 4\beta_{22}^\circ + 2\beta_{23}^\circ = y_{...} = 72. \tag{95}$$

Likewise, the second equation of (81) relates to $\alpha_1$. Its right-hand side is the sum of all observations that have $\alpha_1$ in their model, namely $y_{1..}$. Its left-hand side is $E(y_{1..})$ with $\mathbf{b}$ replaced by $\mathbf{b}^\circ$. Thus the equation is

$$4\mu^\circ + 4\alpha_1^\circ + \beta_{11}^\circ + 3\beta_{12}^\circ = y_{1..} = 32. \tag{96}$$

Suppose that in a design model $\theta_i$ is the effect (parameter) for the $i$th level of the $\theta$ factor. Let $y_{\theta_i}$ be the total of the observations in this level of this factor. Then, the normal equations are

$$[E(y_{\theta_i}) \text{ with } \mathbf{b} \text{ replaced by } \mathbf{b}^\circ] = y_{\theta_i} \tag{97}$$

with $i$ ranging over all levels of all factors $\theta$ including the solitary level of the $\mu$-factor.

The coefficient of each term in (95) is the number of times that its corresponding parameter occurs in $y_{...}$. For example, the coefficient of $\mu^\circ$ is 12 because $\mu$ occurs 12 times in $y_{...}$, the coefficient of $\alpha_1^\circ$ is 4 because $\alpha_1$ occurs four times in $y_{1..}$ and so on.

Likewise, the term in $\beta_{11}^\circ$ in (96) is $\beta_{11}^\circ$ because $\beta_{11}$ occurs once in $y_{1...}$. The term in $\beta_{12}^\circ$ is $3\beta_{12}^\circ$ because $\beta_{12}$ occurs thrice in $y_{1...}$. In general, the coefficients terms in the normal equations (i.e., the elements of $\mathbf{X}'\mathbf{X}$) are the $n_{ij}$'s of the data, determined as follows.

Equation (97) may be called the $\theta_i$ equation, not only because of its form as shown there but also because of its derivation from the least-square procedure when

differentiating with respect to $\theta_i$. The coefficient of $\varphi_j^\circ$ (corresponding to the parameter $\varphi_j$ in (97)) is as follows:

$$\left.\begin{array}{l}\text{Coefficient of } \varphi_j^\circ \\ \text{in the } \theta_i \text{ equation}\end{array}\right\} = \left\{\begin{array}{l}\text{No. of observations in the} \\ i\text{th level of the } \theta\text{-factor} \\ \text{and the } j\text{th level of the } \varphi\text{-factor}\end{array}\right.$$
$$\equiv n(\theta_i, \varphi_j).$$

For example, (96) is the $\alpha_1$-equation and the coefficient of $\beta_{12}^0$ is $n(\beta_{12}, \alpha_1) = n_{12} = 3$ as shown. These $n$'s are the elements of $\mathbf{X'X}$. The property

$$n(\theta_i, \varphi_j) = n(\varphi_j, \theta_i),$$

arising from the definition of $n(\theta_i, \varphi_j)$ just given, accords with the symmetry of $\mathbf{X'X}$. The fact that

$$n(\mu, \theta_i) = n(\theta_i, \mu) = n(\theta_i, \theta_i) = n_{\theta_i} = \left\{\begin{array}{l}\text{No of observations} \\ \text{in the } i\text{th level} \\ \text{of the } \theta\text{-factor}\end{array}\right.$$

is what leads to $\mathbf{X'X}$ having in its first row, in its first column, and in its diagonal, all of the $n$'s (and their various sums) of the data. This is evident in (81) and will be further apparent in subsequent examples. In addition, partitioning the form shown in (81) helps to identify the location of the $n$'s and their sums in $\mathbf{X'X}$. For the illustrative example considered, the $\mu$-equation is first, followed by two $\alpha$-equations, and then by sets of 2 and 3 $\beta$-equations corresponding to the level of the $\beta$-factor within each level of the $\alpha$-factor. Partitioning $\mathbf{X'X}$ in this manner is always helpful in identifying its elements.

## 8. A FEW COMPUTER OUTPUTS

We consider the data from Table 4.11 in Chapter 4. We compare refineries neglecting the source and consider the processes nested within refineries. The SAS output is as follows.

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **refinery** | 3 | 1 2 3 |
| **process** | 2 | 1 2 |
| **Number of observations read** | | 25 |
| **Number of observations used** | | 25 |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: percentage**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 334.010000 | 66.802000 | 1.17 | 0.3588 |
| **Error** | 19 | 1082.550000 | 56.976316 | | |
| **Corrected Total** | 24 | 1416.560000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **percentage Mean** | |
| 0.235790 | 20.26924 | | 7.548266 | 37.24000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **refinery** | 2 | 20.9627778 | 10.4813889 | 0.18 | 0.8334 |
| **process (refinery)** | 3 | 313.0472222 | 104.3490741 | 1.83 | 0.1757 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **refinery** | 2 | 89.3935579 | 44.6967790 | 0.78 | 0.4706 |
| **process (refinery)** | 3 | 313.0472222 | 104.3490741 | 1.83 | 0.1757 |

The code used to generate this output was

```
data efficiency;
input refinery process percentage;
cards;
1 1 31
1 1 33
.......
3 2 37
3 2 43
proc glm;
class refinery process;
model percentage=refinery process(refinery);
run;
```

Galveston, Newark, and Savannah are denoted by 1, 2, and 3, respectively. The source is denoted by 1, and 2.

Note that neither factor was significant.

The R output is

```
Analysis of Variance Table
Response: percent
               Df  Sum Sq Mean Sq F value Pr(>F)
refine          2   20.96  10.481  0.1840 0.8334
refine:process  3  313.05 104.349  1.8314 0.1757
Residuals      19 1082.55  56.976
```

The code used to generate this output was

```
percent<-
c(31,33,44,36,38,26,37,59,42,42,34,42,28,39,36,32,38,42,36,
22,42,46,26,37,43)
> refinery<-c(rep("g",9),rep("n",8),rep("s",8))
> process<-c(1,1,1,1,1,1,2,2,2,1,1,1,1,2,2,2,2,1,1,1,2,2,2,2,2)
> res1 lm(percent~refinery/process)
> anova(res1)
```

## 9. EXERCISES

**1** Suppose that the population of a community consists of 12% who did not complete high school and 68% who did, with the remaining 20% having graduated from college.

Using the data of Table 6.1, find

**(a)** the estimated population average index;

**(b)** the estimated variance of the estimator in (a);

**(c)** the 95% symmetric confidence interval on the population average;

**(d)** the $F$-statistic for testing the hypothesis $H$: $\mu + \alpha_1 = 70$ and $\alpha_1 = \alpha_3 - 15$;

**(e)** a contrast that is orthogonal to $4\alpha_1 - 3\alpha_2 - \alpha_3$;

**(f)** test the hypothesis that the contrast obtained in (e) and $4\alpha_1 - 3\alpha_2 - \alpha_3$ are zero.

**(g)** find 95% simultaneous confidence intervals on the contrast found in (e) and $4\alpha_1 - 3\alpha_2 - \alpha_3$ using

**(i)** the Bonferonni method;

**(ii)** the Scheffe method.

**2** An opinion poll yields the scores of each of the following for some attribute:

**(i)** four laborers as 37, 25, 42, and 28;

**(ii)** two artisans as 23 and 29;

**(iii)** three professionals as 38, 30, and 25; and

**(iv)** two self-employed people as 23 and 29.

For the population from which these people come, the percentages in the four groups are respectively 10%, 20%, 40%, and 30%. What are the estimates and the estimated variances of each of the following?

**(a)** the population score?

**(b)** the difference in score between professionals and an average of the other three groups?

**(c)** the difference between a self-employed and a professional?

**3** (Exercise 2 continued). Test the following hypothesis.

A laborer's score equals an artisan's score equals the arithmetic average of a professional's and a self-employed's score.

**4** (Exercise 2 continued). (a) Find two mutually orthogonal contrasts (one not involving self-employed people) that are orthogonal to the difference between a laborer's and an artisan's score.

**5** (Exercise 2 continued) Suppose that we have yet another professional group with scores 14, 16, 18, 21, 25, and 14. Is the mean score of this group the same as the average of the scores of the other four groups? Perform an appropriate test of the hypothesis.

**6** Suppose that the data of a student opinion poll similar to that of Section 6 of this chapter are as shown below. (Each column represents a section of a course and sections are nested within subjects.)

| English | | | Geology | | Chemistry | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 7 | 8 | 2 | 10 | 8 | 6 | 1 | 8 |
| 5 | 9 | 4 | 6 | 8 | | 2 | 3 | 6 |
| 2 | | 3 | | 9 | | 3 | 2 | |
| | | 4 | | | | 1 | | |

**(a)** Write down the normal equations and find a solution to them.

**(b)** Calculate an analysis of variance table similar to Table 6.7.

**(c)** Test the following hypotheses, one at a time.

**(i)** Sections within courses have the same opinions.

**(ii)** Courses, ignoring sections, have similar opinions.

**(d)** Formulate and test the hypotheses below both simultaneously and independently.

**(i)** Geology's opinion is the mean of English and Chemistry.

**(ii)** English's opinion equals Chemistry both simultaneously and independently. [*Hint:* Do this for the one-way classification model, that is, set all of the $\gamma$'s equal to zero.]

**(e)** Test independently and simultaneously that (section $i$ is the $i$th column):

**(i)** Sections 1 and 3 for English have the same opinion.

**(ii)** Sections 2 and 4 for Chemistry have the same opinion.

**(f)** Find Bonferonni and Scheffe simultaneous 95% confidence intervals on the contrasts in (d) and (e). For the contrasts in (d), use the one-way model. Are the results consistent with those of the hypothesis tests?

**7** For Exercise 6, make a rank transformation and do analysis of variance on the ranks. Compare your results to those in Exercise 6.

**8** Wilson (1993) presents several measurements of the maximum hourly concentrations (in $\mu g/m^3$) of sulfur dioxide ($SO_2$) for each of four power plants. The results with two outliers deleted are as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| Plant 1: | 438 | 619 | 732 | 638 | | |
| Plant 2: | 857 | 1014 | 1153 | 883 | 1053 | |
| Plant 3: | 925 | 786 | 1179 | 786 | | |
| Plant 4: | 893 | 891 | 917 | 695 | 675 | 595 |

**(a)** Perform an analysis of variance to see if there is a significant difference in sulfur dioxide concentration amongst the four plants.

**(b)** Test the hypothesis $H : \alpha_1 - \alpha_4 = 0, \alpha_2 - \alpha_3 = 0, \alpha_1 + \alpha_4 - \alpha_2 - \alpha_3 = 0$.

**(c)** For the contrasts in (b), find:

**(i)** a Bonferonni simultaneous 97% confidence interval.

**(ii)** a Scheffe 99% confidence interval.

**9** Karanthanasis and Pils (2005) present pH measurements of soil specimens taken from three different types of soil. Some measurements are as follow:

| Soil Type | pH Measurements | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alluimum | 6.53 | 6.03 | 6.75 | 6.82 | 6.24 | | | | | | | |
| Glacial Till | 6.07 | 6.07 | 5.36 | 5.57 | 5.48 | 5.27 | 5.80 | 5.03 | 6.65 | | | |
| Residuum | 6.03 | 6.16 | 6.63 | 6.13 | 6.05 | 5.68 | 6.25 | 5.43 | 6.46 | 6.91 | 5.75 | 6.53 |

Determine whether there is a difference in the average pH of the three soil types by performing the Kruskal–Wallis test.

**10** In the model $y_{ij} = \mu_i + e_{ij}, i = 1, ..., a, j = 1, ..., n_i$, show that $\mu_i$ is estimable and find its b.l.u.e.

**11** Consider $n$ observations $y_i$ from a normal population that are serially correlated. This means that $E(y_i) = \mu$ and $var(y_i) = \sigma^2$, the correlation coefficient if $y_i$ and $y_{i+1}$ is $\rho$ for $i = 1, \ldots, n-1$ and that all other correlation coefficients are zero. Show that

**(a)** $E(\bar{y}) = \mu$,

**(b)** $var(\bar{y}) = \frac{\sigma^2}{n}\left[1 + 2\rho\left(1 - \frac{1}{n}\right)\right]$

**(c)** $E(s^2) = \sigma^2\left(1 - \frac{2\rho}{n}\right)$

**12** Derive:

**(a)** The expression for a non-symmetric $(1-\alpha)\,\%$ confidence interval for the contrast $\sum_{i=1}^{a} \lambda_i \alpha_i$ of the one-way classification.

**(b)** The expression for the symmetric confidence interval.

**13** **(a)** Suppose $\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{Z}\mathbf{b}_2 + \mathbf{e}$ with $\mathbf{y} \sim N(\mathbf{X}\mathbf{b}_1 + \mathbf{Z}\mathbf{b}_2, \sigma^2\mathbf{I})$ and that $R(\mathbf{b}_1, \mathbf{b}_2)$ is the reduction in the sum of squares in fitting this model. Prove that $R(\mathbf{b}_2|\mathbf{b}_1)/\sigma^2$ has a non-central $\chi^2$-distribution independent of $R(\mathbf{b}_1)$ and of SSE.

**(b)** Show that $R(\alpha|\mu)/\sigma^2$ and $R(\beta : \alpha|\alpha, \mu)/\sigma^2$ are independently distributed as non-central $\chi^2$ random variables.

**14** Consider the one-way classification model with three treatments. For the tests of hypothesis $H$: all alphas equal show that the numerator of the $F$ statistic is

$$Q = \sum_{i=1}^{3} n_i(\bar{y}_{i.}^2 - \bar{y}_{..})^2$$

**15** Show that the two forms of the Kruskal–Wallis statistic in (73) and (74) are indeed equivalent.

# 7

# THE TWO-WAY CROSSED CLASSIFICATION

This chapter continues with the applications of Chapter 5 that were started in Chapter 6. It will deal at length with the two-way crossed classification (with and without interaction).

## 1. THE TWO-WAY CLASSIFICATION WITHOUT INTERACTION

A course in home economics might include in its laboratory exercises an experiment to illustrate the cooking speed of three makes of pan used with four brands of stove. The students use pans of uniform diameter that are made by different manufacturers to collect data on the number of seconds, beyond three minutes, that it takes to bring two quarts of water to a boil. The experiment is designed to use each of the three makes of pan with each of the four stoves. However, one student carelessly fails to record three of her observations. Her resulting data are shown in Table 7.1. The data include totals for each brand of stove and make of each pan, the number of readings for each, and their mean time. As we have done before, we show the number of readings in parenthesis to distinguish them from the readings themselves.

The observations that the student failed to record are in some sense, "missing observations." We could, if we wanted to, analyze the data using one of the many available "missing observations" techniques available in many books on design of

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

**TABLE 7.1 Number of Seconds (Beyond 3 Minutes) Taken to Boil 2 Quarts of Water**

| | Make of Pan | | | | | |
|---|---|---|---|---|---|---|
| Brand of Stove | A | B | C | Total | No. of Observations | Mean |
| X | 18 | 12 | 24 | 54 | (3) | 18 |
| Y | – | – | 9 | 9 | (1) | 9 |
| Z | 3 | – | 15 | 18 | (2) | 9 |
| W | 6 | 3 | 18 | 27 | (3) | 9 |
| Total | 27 | 15 | 66 | 108 | | |
| No. of observations | (3) | (2) | (4) | | (9) | |
| Mean | 9 | 7.5 | 16.5 | | | 12 |

experiments (e.g., see p. 133 of Federer (1955), or pp. 131–132 of Montgomery (2005)). Most of these techniques involve:

1. estimating the missing observations in some way;
2. putting these estimates into the data;
3. proceeding as if the data were balanced, except for minor adjustments in the degrees of freedom.

We can recommend such procedures on many occasions (see Section 2 of Chapter 8). However, they are of greatest use only when very few observations are missing. This might be the case with Table 7.1, even though 25% of the data have been lost. The data will serve to illustrate techniques for cases where the "missing observations" concept is wholly inappropriate. These include data sets where large numbers of cells may be empty, not because observations were lost but because none were available. Data of this kind occur quite frequently (e.g., Table 4.1). We turn our attention to the analysis of such data using Table 7.1 as an illustration.

The data of Table 7.1 come from a two-way crossed classification. There are two factors with every level of one occurring in combination with every level of the other. We considered models of such data in Section 3 of Chapter 4, paying particular attention to the inclusion of interaction effects in the model. However, it was also pointed out there that, when there is only one observation per cell, the usual model with interactions could not be used.

This is also true of the data in Table 7.1 where there are some cells that do not even have one observation but are empty.

## a. Model

A suitable equation of the model for analyzing the data of Table 7.1 is therefore,

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}. \tag{1}$$

The $y_{ij}$ is the observation of the $i$th row (brand of stove) and $j$th column (make of pan). The mean is $\mu$. The effect of the $i$th row is $\alpha_i$. The effect of the $j$th column is $\beta_j$.

The error term is $e_{ij}$. Equivalently,$\alpha_i$ is the effect due to the $i$th level of the $\alpha$-factor and $\beta_j$ is the effect of the $j$th level of the $\beta$-factor. In general, we have $a$ levels of the $\alpha$-factor with $i = 1, 2, \ldots, a$ and $b$ levels of the $\beta$-factor with $j = 1, 2, \ldots, b$. In the example, $a = 4$ and $b = 3$.

For balanced data, every one of the $ab$ cells of a table like Table 7.1 would have one observation or $n > 1$ observations. The only symbol needed to describe the number of observations in each cell would be ($n \geq 1$). However, in Table 7.1, some cells have zero observations and some have one. Therefore, we need $n_{ij}$ to denote the number of observations in the $i$th row and the $j$th column. In Table 7.1, all $n_{ij} = 0$ or 1. The numbers of observations shown in that table are then the values of

$$n_{i.} = \sum_{j=1}^{b} n_{ij}, n_{.j} = \sum_{i=1}^{a} n_{ij} \quad \text{and} \quad N = n.. = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}.$$

Table 7.1 also shows the corresponding totals and means of the observations. In the next section, we shall also use this convenient $n_{ij}$ notation for data where there are none, one or many observations per cell.

Equation (2) shows the equations of the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ for the observations in Table 7.1. We show the elements of $\mathbf{b}$, namely $\mu, \alpha_1, \ldots, \alpha_4, \beta_1, \beta_2,$ and $\beta_3$ both as a vector and as headings to the columns of the matrix $\mathbf{X}$. This is purely for convenience in reading the equations. It clarifies the incidence of the elements in the model, as does the partitioning, according to the different factors $\mu, \alpha,$ and $\beta$. The model equations for the data in Table 7.1 are

$$\begin{array}{c} \\ \begin{bmatrix} 18 \\ 12 \\ 24 \\ 9 \\ 3 \\ 15 \\ 6 \\ 3 \\ 18 \end{bmatrix} \end{array} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{23} \\ y_{31} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \end{bmatrix} = \begin{array}{c} \begin{matrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \beta_1 & \beta_2 & \beta_3 \end{matrix} \\ \left[\begin{array}{c:cccc:ccc} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{23} \\ e_{31} \\ e_{33} \\ e_{41} \\ e_{42} \\ e_{43} \end{bmatrix}. \tag{2}$$

### b. Normal Equations

For the given $\mathbf{X}$ and observations $\mathbf{y}$, Equations (2) are in the form of $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$. We can write the corresponding normal equations $\mathbf{X'Xb}^{\circ} = \mathbf{X'y}$, in a manner similar to (2). They are

$$\begin{array}{c} \mu^\circ\ \alpha_1^\circ\ \alpha_2^\circ\ \alpha_3^\circ\ \alpha_4^\circ\ \beta_1^\circ\ \beta_2^\circ\ \beta_3^\circ \\ \left[\begin{array}{c:cccc:ccc} 9 & 3 & 1 & 2 & 3 & 3 & 2 & 4 \\ \hdashline 3 & 3 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 3 & 0 & 0 & 0 & 3 & 1 & 1 & 1 \\ \hdashline 3 & 1 & 0 & 1 & 1 & 3 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 & 2 & 0 \\ 4 & 1 & 1 & 1 & 1 & 0 & 0 & 4 \end{array}\right] \end{array} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \alpha_4^\circ \\ \beta_1^\circ \\ \beta_2^\circ \\ \beta_3^\circ \end{bmatrix} = \begin{bmatrix} 108 \\ 54 \\ 9 \\ 18 \\ 27 \\ 27 \\ 15 \\ 86 \end{bmatrix}. \tag{3}$$

We gave some general properties of these equations in Section 4 of Chapter 6. These are further evident here. In this case, the first row and column and the diagonal of $\mathbf{X}'\mathbf{X}$ have $n_{..}$, the $n_{i.}$'s, and the $n_{.j}$'s in them. The only other non-zero off-diagonal elements are those in the $a \times b$ matrix of 1's and 0's (and its transpose) corresponding to the pattern of observations. The partitioning indicated in (3) highlights the form of $\mathbf{X}'\mathbf{X}$ and suggests how we would accommodate more levels of the factor.

### c. Solving the Normal Equations

In the examples of Sections 2 and 6 of Chapter 6, solutions of the normal equations were easily derived by the procedure indicated in equation (4) of Chapter 6. Now, however, even after making use of that procedure, there is no neat explicit solution. We can obtain a numerical solution but algebraically we cannot represent it succinctly.

In (3), the sum of the $a$ rows of $\mathbf{X}'\mathbf{X}$ immediately after the first (the $\alpha$-equations) equals the first row. The sum of the last $b$ rows (the $\beta$-equations) also equals the first row. Since $\mathbf{X}'\mathbf{X}$ has order $q = 1 + a + b$, its rank is $r = r(\mathbf{X}'\mathbf{X}) = 1 + a + b - 2 = a + b - 1$. Thus $p - r = 2$. We may solve (3) by setting an appropriate two elements of $\mathbf{b}^\circ$ equal to zero and deleting the corresponding equations. One of the easiest ways to do this is to put $\mu^\circ = 0$ and either $\alpha_1^\circ = 0$ or $\beta_b^\circ = 0$, according to whether $a < b$ or $a > b$. When $a = b$, it is immaterial. Thus, when there are fewer $\alpha$-levels than $\beta$-levels, put $\alpha_1^\circ = 0$ and when there are fewer $\beta$-levels than $\alpha$-levels, put $\beta_b^\circ = 0$. In our example, there are fewer $\beta$-levels than $\alpha$-levels. Thus with $\mu^\circ = 0 = \beta_3^\circ$, we get from (3),

$$\left[\begin{array}{cccc:cc} 3 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 & 1 & 1 \\ \hdashline 1 & 0 & 1 & 1 & 3 & 0 \\ 1 & 0 & 0 & 1 & 0 & 2 \end{array}\right] \begin{bmatrix} \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \alpha_4^\circ \\ \beta_1^\circ \\ \beta_2^\circ \end{bmatrix} = \begin{bmatrix} y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{4.} \\ y_{.1} \\ y_{.2} \end{bmatrix} = \begin{bmatrix} 54 \\ 9 \\ 18 \\ 27 \\ 27 \\ 15 \end{bmatrix}. \tag{4}$$

Written in full, these equations are

$$\begin{aligned} 3\alpha_1^\circ \qquad\qquad + \beta_1^\circ + \beta_2^\circ &= 54 \\ \alpha_2^\circ \qquad\qquad\qquad &= 9 \\ 2\alpha_3^\circ \qquad + \beta_1^\circ \qquad &= 18 \\ 3\alpha_4^\circ + \beta_1^\circ + \beta_2^\circ &= 27 \end{aligned} \tag{5}$$

and

$$\begin{aligned} \alpha_1^\circ \quad +\alpha_3^\circ + \alpha_4^\circ + 3\beta_1^\circ \qquad &= 27 \\ \alpha_1^\circ \qquad\quad +\alpha_4^\circ \qquad 2\beta_2^\circ &= 15. \end{aligned} \tag{6}$$

From (5), the $\alpha^\circ$'s are expressed in terms of the $\beta^\circ$'s. Substitution in (6) then leads to the solutions for the $\beta^\circ$'s. Thus (5) gives

$$\begin{aligned} \alpha_1^\circ &= 54/3 - \tfrac{1}{3}(\beta_1^\circ + \beta_2^\circ) = 18 - \tfrac{1}{3}[1(\beta_1^\circ) + 1(\beta_2^\circ)] \\ \alpha_2^\circ &= 9/1 \qquad\qquad\quad = 9 - \tfrac{1}{1}[0(\beta_1^\circ) + 0(\beta_2^\circ)] \\ \alpha_3^\circ &= 18/2 - \tfrac{1}{2}\beta_1^\circ \qquad = 9 - \tfrac{1}{2}[1(\beta_1^\circ) + 0(\beta_2^\circ)] \\ \alpha_4^\circ &= 27/3 - \tfrac{1}{3}(\beta_1^\circ + \beta_2^\circ) = 9 - \tfrac{1}{3}[1(\beta_1^\circ) + 1(\beta_2^\circ)]. \end{aligned} \tag{7}$$

The reason for including the coefficients 1 and 0 on the right-hand sides of (5) will become clear when we consider the generalization of the procedure. For this reason, we retain the 1's and the 0's. Substituting (7) into (6) gives

$$\begin{aligned} &\{3 - [1(1)/3 + 0(0)/1 + 1(1)/2 + 1(1)/3]\}\beta_1^\circ \\ &\quad -[1(1)/3 + 0(0)/1 + 1(0)/2 + 1(1)/3]\beta_2^\circ \\ &\quad = 27 - [1(18) + 0(9) + 1(9) + 1(9)]. \end{aligned} \tag{8}$$

$$\begin{aligned} &-[1(1)/3 + 0(0)/1 + 0(1)/2 + 1(1)/3]\beta_1^\circ \\ &\quad +\{2 - [1(1)/3 + 0(0)/1 + 0(0)/2 + 1(1)/3]\}\beta_2^\circ . \\ &\quad = 15 - [1(18) + 0(9) + 0(9) + 1(9)] \end{aligned}$$

Equations (8) reduce to

$$(11/6)\beta_1^\circ - (4/6)\beta_2^\circ = -9 \quad \text{and} \quad (-4/6)\beta_1^\circ + (8/6)\beta_2^\circ = -12. \tag{9}$$

The solutions to Equations (8) are

$$\beta^\circ = \begin{bmatrix} \beta_1^\circ \\ \beta_2^\circ \end{bmatrix} = \begin{bmatrix} -10 \\ -14 \end{bmatrix}. \tag{10}$$

Substituting the values obtained in (10) into (7), we get

$$\alpha_1^\circ = 26, \quad \alpha_2^\circ = 9, \quad \alpha_3^\circ = 14, \quad \text{and} \quad \alpha_4^\circ = 17.$$

The resulting solution to the normal equations is

$$\mathbf{b}^{\circ\prime} = \begin{bmatrix} 0 & 26 & 9 & 14 & 17 & -10 & -14 & 0 \end{bmatrix}. \tag{11}$$

### d. Absorbing Equations

Development of (11) as a solution to (3) illustrates what is sometimes called the *absorption process.* This is because in going from (4) to (8), the $\alpha$-equations of (5) are "absorbed" into the $\beta$-equations of (6). Here, we see the reason given in Sub-section c above for the rule about deciding whether to put $\alpha_1^\circ = 0$ or $\beta_b^\circ = 0$. The objective is for (8) to have as few equations as possible. Hence, if there are fewer $\beta$-levels than $\alpha$-levels, we put $\beta_b^\circ = 0$, absorb the $\alpha$-equations and have equations (8) in terms of $(b-1)\beta^\circ$'s. On the other hand, if $a < b$, we put $\alpha_1^\circ = 0$, absorb the $\beta$-equations, and have equations like (8) in terms of $(a-1)\alpha$'s. It is of no consequence in using the ultimate solution, which one is obtained. The important thing is the number of equations in (8), either $a-1$ or $b-1$, whichever is less. In many instances, the number of equations is, in fact, of little importance because, even if one of $a$ and $b$ is much larger than the other, the solution of (8) will require a computer. However, in Chapter 9, we discuss situations in which one of $a$ or $b$ is considerably larger than the other ($a = 10$ and $b = 2000$, say), and then the method of obtaining (8) is of material importance.

We now describe the absorption process in general terms. Akin to (3), the normal equations are

$$\left[\begin{array}{c|ccc|ccc} n_{\cdot\cdot} & n_{1\cdot} & \cdots & n_{a\cdot} & n_{\cdot 1} & \cdots & n_{\cdot a} \\ \hline n_{1\cdot} & n_{1\cdot} & & \mathbf{0} & & & \\ \vdots & & \ddots & & & \{n_{ij}\} & \\ n_{a\cdot} & \mathbf{0} & & n_{a\cdot} & & & \\ \hline n_{\cdot 1} & & & & n_{\cdot 1} & & \mathbf{0} \\ \vdots & & \{n_{ij}\} & & & \ddots & \\ n_{\cdot a} & & & & \mathbf{0} & & n_{\cdot b} \end{array}\right]\left[\begin{array}{c} \mu^\circ \\ \hline \alpha_1^\circ \\ \vdots \\ \alpha_a^\circ \\ \hline \beta_1^\circ \\ \vdots \\ \beta_b^\circ \end{array}\right] = \left[\begin{array}{c} y_{\cdot\cdot} \\ \hline y_{1\cdot} \\ \vdots \\ y_{a\cdot} \\ \hline y_{\cdot 1} \\ \vdots \\ y_{\cdot a} \end{array}\right] \tag{12}$$

Analogous to (4), if we put $\mu^\circ = 0$ and $\beta_b^\circ = 0$, equations (12) reduce to

$$\left[\begin{array}{ccc|ccc} n_{1\cdot} & & \mathbf{0} & n_{11} & \cdots & n_{1,b-1} \\ & \ddots & & \vdots & & \vdots \\ \mathbf{0} & & n_{a\cdot} & n_{a1} & \cdots & n_{a,b-1} \\ \hline n_{11} & \cdots & n_{a1} & n_{\cdot 1} & & \mathbf{0} \\ \vdots & & \vdots & & \ddots & \\ n_{1,b-1} & \cdots & n_{a,b-1} & \mathbf{0} & & n_{\cdot,b-1} \end{array}\right]\left[\begin{array}{c} \alpha_1^\circ \\ \vdots \\ \alpha_a^\circ \\ \hline \beta_1^\circ \\ \vdots \\ \beta_{b-1}^\circ \end{array}\right] = \left[\begin{array}{c} y_{1\cdot} \\ \vdots \\ y_{a\cdot} \\ \hline y_{\cdot 1} \\ \vdots \\ y_{\cdot,b-1} \end{array}\right]. \tag{13}$$

Solving the first $a$ equations of (13) gives

$$\alpha_i^\circ = \bar{y}_{i.} - \frac{1}{n_{i.}} \sum_{j=1}^{b-1} n_{ij}\beta_j^\circ \quad \text{for} \quad i = 1, 2, \ldots, a, \tag{14}$$

as in (7). Substitution of these values of $\alpha_i^\circ$ in the last $b-1$ equations of (13) gives

$$\left( n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}} \right) \beta_j^\circ - \sum_{j \neq j'}^{b-1} \left( \sum_{i=1}^{a} \frac{n_{ij} n_{ij'}}{n_{i.}} \right) \beta_{j'}^\circ = y_{.j} - \sum_{i=1}^{a} n_{ij} \bar{y}_{i.} \tag{15}$$

$$\text{for } j, j' = 1, 2, \ldots, b-1$$

$$\mathbf{C}\boldsymbol{\beta}_{b-1}^\circ = \mathbf{r} \quad \text{with solution} \quad \boldsymbol{\beta}_{b-1}^\circ = \mathbf{C}^{-1}\mathbf{r} \tag{16}$$

where,

$$\mathbf{C} = \{c_{jj'}\} \quad \text{and} \quad \mathbf{r} = \{r_j\} \quad \text{for} \quad j = 1, \ldots, b-1$$

with

$$c_{jj} = n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}}, \quad c_{jj'} = -\sum_{i=1}^{a} \frac{n_{ij} n_{ij'}}{n_{i.}} \quad \text{for} \quad j \neq j' \tag{17}$$

and

$$r_j = y_{.j} - \sum_{i=1}^{a} n_{ij} \bar{y}_{i.} \quad \text{for} \quad j = 1, \ldots, b-1. \tag{18}$$

We can check these calculations by also calculating $c_{bb}, c_{jb}, \text{and} r_b$ and confirming that

$$\sum_{j'=1}^{b} c_{jj'} = 0 \quad \text{for all } j, \quad \text{and} \quad \sum_{j=1}^{b} r_j = 0.$$

The solution $\boldsymbol{\beta}_{b-1}^\circ$ in (16) is subscripted to emphasize that it has $b-1$ and not $b$ elements. To express the solutions $\alpha_i^\circ$ in matrix form, we write

$$\alpha^\circ = \begin{bmatrix} \alpha_1^\circ \\ \vdots \\ \alpha_a^\circ \end{bmatrix}, \quad \mathbf{y}_a = \begin{bmatrix} y_{1.} \\ \vdots \\ y_{a.} \end{bmatrix},$$

and $\mathbf{D}_\alpha = \mathbf{D}\{n_{i.}\}$, for $i = 1, 2, \ldots, a$, a diagonal matrix (See Section 1 of Chapter 1) of order $a$ of the $n_{i.}$ values. We also define

$$\mathbf{N}_{a\times(b-1)} = \begin{bmatrix} n_{11} & \cdots & n_{1,b-1} \\ \vdots & & \vdots \\ n_{a1} & \cdots & n_{a,b-1} \end{bmatrix},$$

$$\mathbf{M}_{a\times(b-1)} = \mathbf{D}_a^{-1}\mathbf{N} = \left\{ \frac{n_{ij}}{n_{i.}} \right\} \quad \text{for} \quad i = 1, \ldots, a \quad \text{and} \quad j = 1, \ldots, b-1,$$

and

$$\bar{\mathbf{y}}_a = \mathbf{D}_a^{-1}\mathbf{y}_a = \{\bar{y}_{i.}\} \quad \text{for} \quad i = 1, \ldots, a. \tag{19}$$

Then, from (13),

$$\boldsymbol{\alpha}^\circ = \mathbf{D}_a^{-1} - \mathbf{M}\boldsymbol{\beta}_{b-1}^\circ = \bar{\mathbf{y}}_a - \mathbf{M}\boldsymbol{\beta}_{b-1}^\circ.$$

Thus,

$$\mathbf{b}^\circ = \begin{bmatrix} 0 \\ \boldsymbol{\alpha}^\circ \\ \boldsymbol{\beta}_{b-1}^\circ \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{\mathbf{y}}_a - \mathbf{M}\mathbf{C}^{-1}\mathbf{r} \\ \mathbf{C}^{-1}\mathbf{r} \\ 0 \end{bmatrix}. \tag{20}$$

Section 4 of this chapter deals with the condition of "connectedness" of unbalanced data. Although most modestly sized sets of data are usually connected, large sets of survey-style data are sometimes not connected. The condition of connectedness is important because only when data are connected do $\mathbf{C}^{-1}$ and the solution in (20) exist. Further discussion therefore relates solely to data that are connected. This condition must be satisfied before we can undertake this analysis. Section 4 indicates how to ascertain if data are connected.

Corresponding to the solution (20), the generalized inverse of $\mathbf{X}'\mathbf{X}$ of (12) is

$$\mathbf{G} = \begin{bmatrix} 0 & \mathbf{0} & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{D}_a^{-1} + \mathbf{M}\mathbf{C}^{-1}\mathbf{M}' & -\mathbf{M}\mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{C}^{-1}\mathbf{M}' & \mathbf{C}^{-1} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{0} & 0 \end{bmatrix}. \tag{21}$$

The non-null part of the matrix is of course the regular inverse of the matrix of coefficients in equations (13). Thus, $\mathbf{G}$ is in accord with Section 7 of Chapter 5.

**Example 1 Generalized Inverse of X′X in (4)** From (9),

$$\mathbf{C}^{-1} = \begin{bmatrix} \frac{11}{6} & -\frac{4}{6} \\ -\frac{4}{6} & \frac{8}{6} \end{bmatrix}^{-1} = \frac{1}{12}\begin{bmatrix} 8 & 4 \\ 4 & 11 \end{bmatrix}.$$

From (4),

$$\mathbf{D}_a = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

and

$$\mathbf{N} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad \text{so that} \quad \mathbf{M} = \mathbf{D}_a^{-1}\mathbf{N} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ 0 & 0 \\ \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Therefore, for use in **G**,

$$\mathbf{MC}^{-1} = \frac{1}{12}\begin{bmatrix} 4 & 5 \\ 0 & 0 \\ 4 & 2 \\ 4 & 5 \end{bmatrix} \quad \text{and} \quad \mathbf{MC}^{-1}\mathbf{M}' = \frac{1}{12}\begin{bmatrix} 3 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 2 \\ 3 & 0 & 2 & 3 \end{bmatrix}.$$

Substitution of the various sub-matrices into (21) gives

$$\mathbf{G} = \frac{1}{12}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7 & 0 & 2 & 3 & -4 & -5 & 0 \\ 0 & 0 & 12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 8 & 2 & -4 & -2 & 0 \\ 0 & 3 & 0 & 2 & 7 & -4 & -5 & 0 \\ 0 & -4 & 0 & -4 & -4 & 8 & 4 & 0 \\ 0 & -5 & 0 & -2 & -5 & 4 & 11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Post-multiplication of the right-hand side of equation (3) gives the solution of $\mathbf{GX}'\mathbf{y} = \mathbf{b}^{\circ}$ as shown in (11). □

### e. Analyses of Variance

(*i*) ***Basic Calculations.*** The reduction of sum of squares is $R(\mu, \alpha, \beta) = \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$. In preceding examples, $\mathbf{b}^{\circ}\mathbf{X}'\mathbf{y}$ is simplified. The simplification is not possible here because of the way $\mathbf{b}^{\circ}$ has been derived. However, if we define

$$\mathbf{y}'_{\beta} = \begin{bmatrix} y_{.1} & \cdots & y_{.,b-1} \end{bmatrix}, \tag{22}$$

then from (20), we find that

$$R(\mu, \alpha, \beta) = (\bar{y}_a - \mathbf{M}\mathbf{C}^{-1}\mathbf{r})'\mathbf{y}_a + (\mathbf{C}^{-1}\mathbf{r})'\mathbf{y}_{\beta}.$$

However, since it follows from (18) that

$$\mathbf{r} = \mathbf{y}_{\beta} - \mathbf{M}'\mathbf{y}_a,$$

and the expression for $R(\mu, \alpha, \beta)$ further simplifies to

$$R(\mu, \alpha, \beta) = \bar{\mathbf{y}}'_a\mathbf{y}_a + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}. \tag{23}$$

As usual, we have

$$R(\boldsymbol{\mu}) = n_{\cdot\cdot}\bar{y}_{\cdot\cdot}^2 = \frac{y_{\cdot\cdot}^2}{n_{\cdot\cdot}}. \tag{24}$$

In line with Section 3a of Chapter 6, using (19),

$$R(\mu, \alpha) = \sum_{i=1}^{a} n_{i\cdot}\bar{y}_{i\cdot}^2 = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_{i\cdot}} = \bar{\mathbf{y}}'_a\bar{\mathbf{y}}_a. \tag{25}$$

Hence, in (23),

$$\begin{aligned} R(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= R(\boldsymbol{\mu}, \boldsymbol{\alpha}) + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} \\ &= \sum_{i=1}^{a} n_{i\cdot}\bar{y}_{i\cdot}^2 + \boldsymbol{\beta}^{\circ\prime}\mathbf{r} \end{aligned} \tag{26}$$

with the terms of $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r} = \boldsymbol{\beta}^{\circ\prime}\mathbf{r}$ defined in (16), (17), and (18).

We now calculate the terms in (24), (25), and (26) for the data in Table 7.1. The results are

$$R(\mu) = \frac{108^2}{9} = 1296, \tag{27}$$

$$R(\mu, \alpha) = \frac{54^2}{3} + \frac{9^2}{1} + \frac{18^2}{2} + \frac{27^2}{3} = 1458 \tag{28}$$

and, using (10) for $\boldsymbol{\beta}^{\circ\prime}$ and (9) for $\mathbf{r}$,

$$R(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 1458 + (-10)(-9) + (-14)(-12) = 1716. \tag{29}$$

**(*ii*) *Fitting the Model.*** The first analysis of variance that we shall consider is that for fitting the model. This partitions $R(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, the sum of squares for fitting the model in two parts. They are $R(\mu)$ for fitting the mean and $R(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{\mu})$ for fitting the $\alpha$- and $\beta$-factors after the mean. The latter is

$$\begin{aligned} R(\alpha, \beta|\mu) &= R(\mu, \alpha, \beta) - R(\mu) \\ &= \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} - N\bar{y}_{..}^2 \end{aligned} \tag{30}$$

from (24) and (26). We note what is obvious here.

$$R(\mu) + R(\alpha, \beta|\mu) = R(\mu, \alpha, \beta)$$

by the definition of $R(\alpha, \beta|\mu)$. The values for the terms for the data of Table 7.1 are $R(\mu) = 1296$ from (27) and

$$R(\alpha, \beta|\mu) = 1716 - 1296 = 420$$

from (27) and (29). These and the other terms of the analysis,

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 = 18^2 + \cdots + 18^2 = 1728$$

and

$$\text{SSE} = \text{SST} - R(\alpha, \beta|\mu) = 1728 - 1716 = 12$$

using (29) are shown in Table 7.2a. Table 7.2a also contains the corresponding $F$-statistics (based on the normality of the $e$'s). These are $F(M) = 324$ and $F(R_{\text{m}}) = 21$. They are significant at the 5% level. The tabulated values of $F_{1,3}$ and $F_{5,3}$ at the 5% level are 10.13 and 9.01, respectively. Therefore, we reject the hypothesis that $E(\bar{y})$ is zero. We further conclude that the model needs something more than just $\mu$ in order to satisfactorily explain variation in the $y$ variable.

**(*iii*) *Fitting Rows Before Columns.*** The significance of the statistic $F(R_{\text{m}})$ in Table 7.2a leads us to enquire whether it is the $\alpha$'s (rows, or brands of stoves), or the $\beta$'s (columns, or makes of pan), or both that are contributing to this significance. First, consider the $\alpha$'s in terms of fitting the model

$$y_{ij} = \mu + \alpha_i + e_{ij}.$$

**TABLE 7.2 Analyses of Variance for Two-Way Classification, No Interaction (Data of Table 7.1)**

**(a) For fitting $\mu$, and $\alpha$ and $\beta$ after $\mu$**

| Source of variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean[a] | $1 = 1$ | $R(\mu) = 1296$ | 1296 | $F(M) = 324$ |
| $\alpha$ and $\beta$ after $\mu$ | $a + b - 2 = 5$ | $R(\alpha,\beta\|\mu) = 420$ | 84 | $F(R_{\mathrm{m}}) = 21$ |
| Residual error | $N' = 3$ | SSE = 12 | 4 | |
| Total | $N = 9$ | SST = 1728 | | |

**(b) For fitting $\mu$, $\alpha$ *after* $\mu$, and $\beta$ after $\mu$ and $\alpha$**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean | $1 = 1$ | $R(\mu) = 1296$ | 1296 | $F(M) = 324$ |
| $\alpha$ after $\mu$ | $a - 1 = 3$ | $R(\alpha\|\mu) = 162$ | 54 | $F(\alpha\|\mu) = 13.5$ |
| $\beta$ after $\mu$ and $\alpha$ | $b - 1 = 2$ | $R(\beta\|\mu,\alpha) = 258$ | 129 | $F(\beta\|\mu,\alpha) = 32.25$ |
| Residual error | $N' = 3$ | SSE = 12 | 4 | |
| Total | $N = 9$ | SST = 1728 | | |

**(c) For fitting $\mu$, $\beta$ after $\mu$, and $\alpha$ after $\mu$ and $\beta$**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| Mean | $1 = 1$ | $R(\mu) = 1296$ | 1296 | $F(M) = 324$ |
| $\beta$ after $\mu$ | $b - 1 = 2$ | $R(\beta\|\mu) = 148.5$ | 74.25 | $F(\beta\|\mu) = 18.5625$ |
| $\alpha$ after $\mu$ and $\beta$ | $a - 1 = 3$ | $R(\alpha\|\mu,\beta) = 271.5$ | 90.5 | $F(\alpha\|\mu,\beta) = 22.625$ |
| Residual error | $N' = 3$ | SSE = 12 | 4 | |
| Total | $N = 9$ | SST = 1728 | | |

[a] $N' = N - a - b + 1$

Since this is just the model for the one-way classification, the sum of squares for fitting it is just $R(\mu, \alpha)$ as given in (25). Therefore, the sum of squares attributable to fitting $\alpha$ after $\mu$ is from (24) and (25).

$$\begin{aligned} R(\alpha|\mu) &= R(\mu, \alpha) - R(\mu) \\ &= \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 - n..\bar{y}_{..}^2. \end{aligned} \tag{31}$$

Furthermore, from (26), the sum of squares attributable to fitting the $\beta$'s after $\mu$ and the $\alpha$'s is

$$\begin{aligned} R(\beta|\mu, \alpha) &= R(\mu, \alpha, \beta) - R(\mu, \alpha) \\ &= \beta^{\circ\prime}\mathbf{r} = \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}. \end{aligned} \tag{32}$$

The sums of squares in (31) and (32) are shown in Table 7.2b. Of course, they are a portioning of $R(\alpha, \beta|\mu)$ shown in Table 7.2a, since

$$\begin{aligned} R(\alpha|\mu) + R(\beta|\mu, \alpha) &= R(\mu, \alpha) - R(\mu) + R(\mu, \alpha, \beta) - R(\mu, \alpha) \\ &= R(\mu, \alpha, \beta) - R(\mu) = R(\alpha, \beta|\mu). \end{aligned} \tag{33}$$

Likewise, all three $R$'s shown in Table 7.2b sum to $R(\mu, \alpha, \beta)$ because

$$R(\mu) + R(\alpha|\mu) + R(\beta|\mu, \alpha) = R(\mu) + R(\alpha, \beta|\mu) = R(\mu, \alpha, \beta). \tag{34}$$

Calculation of $R(\beta|\mu, \alpha)$,$R(\alpha|\mu)$, and $R(\beta|\mu, \alpha)$ for Table 7.2b is as follows. Substituting in (31) from (27) and (28) yields

$$R(\alpha|\mu) = 1458 - 1296 = 162.$$

Substituting in (32) from (9) and (10) gives

$$R(\beta|\mu, \alpha) = -9(-10) - 12(-14) = 258. \tag{35}$$

The validity of (33) follows because

$$R(\alpha|\mu) + R(\beta|\mu, \alpha) = 162 + 258 = 420 = R(\alpha, \beta|\mu) \text{ of Table 7.2a}$$

Table 7.2b shows the $F$-statistics corresponding to the $R$'s. Comparing $F(\alpha|\mu) = 13.5$ and $F(\beta|\mu, \alpha) = 32.25$ to the tabulated values of the $F_{3,3}$ and $F_{2,3}$-distributions, respectively, namely 9.28 and 9.55 at the 5% level, we conclude that having both $\alpha$-effects and $\beta$-effects in the model adds significantly to its adequacy in terms of explaining the variation in $y$.

**(*iv*) *Fitting Columns Before Rows.*** Table 7.2b is for fitting $\mu$, $\mu$ and $\alpha$, and then $\mu$, $\alpha$, and $\beta$. However, we could just as well consider the $\alpha$'s and $\beta$'s in the reverse order and contemplate fitting $\mu$,$\mu$ and $\beta$, and then $\mu$, $\alpha$, and $\beta$. To do this, we would first fit the model

$$y_{ij} = \mu + \beta_j + e_{ij}.$$

This leads to

$$R(\mu, \beta) = \sum_{j=1}^{b} n_{.j}\bar{y}_{.j}^2 \tag{36}$$

similar to (25). Then, analogous to (31), we have

$$\begin{aligned} R(\beta|\mu) &= R(\mu, \beta) - R(\mu) \\ &= \sum_{j=1}^{b} n_{.j}\bar{y}_{.j}^2 - n_{..}\bar{y}_{..}^2. \end{aligned} \tag{37}$$

We also have, similar to the first part of (32),

$$R(\alpha|\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta) \tag{38}$$

for the sum of squares due to fitting the $\alpha$ after fitting $\mu$ and $\beta$. However, we do not have an expression for $R(\alpha|\mu,\beta)$ analogous to $\beta^{\circ\prime}\mathbf{r}$ of (32). By means of (34), we can avoid needing such an expression, because using (34) in (38) gives

$$\begin{aligned} R(\alpha|\mu,\beta) &= R(\mu,\alpha) + R(\beta|\mu,\alpha) - R(\mu,\beta) \\ &= \sum_{i=1}^{a} n_{i\cdot}\bar{y}_{i\cdot}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} - \sum_{j=1}^{b} n_{\cdot j}\bar{y}_{\cdot j}^2 \end{aligned} \tag{39}$$

on substituting from (25), (32), and (36), respectively. Hence, having once obtained $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$, we have $R(\alpha|\mu,\beta)$ directly available without further ado. Of course, analogues of (33) and (34) are also true. We have that

$$R(\beta|\mu) + R(\alpha|\mu,\beta) = R(\alpha,\beta|\mu) \tag{40a}$$

and

$$R(\mu) + R(\beta|\mu) + R(\alpha|\beta,\mu) = R(\mu,\alpha,\beta). \tag{40b}$$

With the data of Table 7.1, equation (36) is

$$R(\mu,\beta) = \frac{27^2}{3} + \frac{15^2}{2} + \frac{66^2}{4} = 1444.5. \tag{41}$$

Using (41) and (27) in (37) gives

$$R(\beta|\mu) = 1444.5 - 1296 = 148.5$$

Then in (39),

$$R(\alpha|\beta,\mu) = 1458 + 258 - 1444.5 = 271.5$$

from (28), (35), and (41), respectively. We note that as indicated in (40),

$$R(\beta|\mu) + R(\alpha|\mu,\beta) = 148.5 + 271.5 = 420 = R(\alpha,\beta|\mu)$$

as shown in Table 7.2a.

The $F$-statistics corresponding to $R(\beta|\mu)$ and $R(\alpha|\mu,\beta)$ in Table 7.2c are both significant at the 5% level. (The tabulated values are 9.55 and 9.28 for comparing $F(\beta|\mu)$ and $F(\alpha|\mu,\beta)$, respectively.) We therefore conclude that including both $\beta$-effects and $\alpha$-effects in the model adds significantly to its interpretive value.

Table 7.2 shows the analyses of variance for the data of Table 7.1. In contrast, Table 7.3 shows the analysis of variance (excluding mean squares and $F$-statistics)

**TABLE 7.3 Analyses of Variance for Two-Way Classification, No Interaction**

**Table 7.3 (a) For fitting $\mu$, and $\alpha$ and $\beta$ after $\mu$**

| Source of Variation | d.f.[a] | Sum of Squares[b] | Equation |
|---|---|---|---|
| Mean, $\mu$ | 1 | $R(\mu) = n\bar{y}_{..}^2$ | (24) |
| $\alpha$ and $\beta$ after $\mu$ | $a + b - 2$ | $R(\alpha, \beta \mid \mu) = \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} - n..\bar{y}_{..}^2$ | (30) |
| Residual error[c] | $N'$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 - \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ | |
| Total | $N$ | $\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2$ | |

**Table 7.3 (b) For fitting $\mu$,$\alpha$ after$\mu$, and $\beta$ after $\mu$ and $\alpha$**

| Source of Variation | d.f.[a] | Sum of Squares | Equation |
|---|---|---|---|
| Mean, $\mu$ | $1 = 1$ | $R(\mu) = n..\bar{y}_{..}^2$ | (24) |
| $\alpha$ after $\mu$ | $a - 1$ | $R(\alpha \mid \mu) = \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 - n..\bar{y}_{..}^2$ | (31) |
| $\beta$ after $\mu$ and $\alpha$ | $b - 1$ | $R(\beta \mid \mu, \alpha) = \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ | (32) |
| Residual error | $N'$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 - \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ | |
| Total | $N$ | $\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2$ | |

**Table 7.3 (c) For fitting $\mu$, $\beta$ after $\mu$, and $\alpha$ after $\mu$ and $\beta$**

| Source of Variation | d.f.[a] | Sum of Squares[b] | Equation |
|---|---|---|---|
| Mean, $\mu$ | 1 | $R(\mu) = n..\bar{y}_{..}^2$ | |
| $\beta$ after $\mu$ | $b - 1$ | $R(\beta \mid \mu) = \sum_{j=1}^{b} n_{.j}\bar{y}_{.j}^2 - n..\bar{y}_{..}^2$ | |
| $\alpha$ after $\beta$ and $\mu$ | $a - 1$ | $R(\alpha \mid \mu, \beta) = \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} - \sum_{j=1}^{b} n_{.j}\bar{y}_{.j}^2$ | |
| Residual error | $N'$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \sum_{i=1}^{a} n_{i.}\bar{y}_{i.}^2 - \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ | |
| Total | $N = 9$ | $\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2$ | |

[a] $N \equiv n$ and $N' = N - a - b + 1$

[b] $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ is obtained from equations (16)–(18)

[c] Summations are for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$.

for the general case. It also shows the equations from which the expressions for the sum of squares have been derived.

**(*v*) *Ignoring and/or Adjusting for Effects.*** In Tables 7.2b and 7.3b, the sums of squares have been described as

$R(\mu)$: due to fitting a mean $\mu$,
$R(\alpha|\mu)$: due to fitting $\alpha$ after $\mu$,

and

$R(\beta|\mu, \alpha)$: due to fitting $\beta$ adjusted for $\mu$ and $\alpha$.

This description carries with it a sequential concept, of first fitting $\mu$, then $\mu$ and $\alpha$, and then $\mu, \alpha$, and $\beta$. An alternative description, similar to that used by some writers, is

$R(\mu)$: due to fitting $\mu$, ignoring $\alpha$ and $\beta$,
$R(\alpha|\mu)$: due to fitting $\alpha$, adjusted for $\mu$ and ignoring $\beta$,

and

$R(\beta|\mu, \alpha)$ due to fitting $\beta$ adjusted for $\mu$ and $\alpha$.

On many occasions, of course, Tables 7.2 and 7.3 are shown without the $R(\mu)$ line, and with the SST line reduced by $R(\mu)$ so that it has $N - 1$ degrees of freedom, and the sum of squares $\mathrm{SST_m} = \mathbf{y}'\mathbf{y} - N\bar{y}_{\cdot\cdot}^2$. In that case, the mention of $\mu$ in the descriptions of $R(\alpha|\mu)$ and $R(\beta|\mu, \alpha)$ is then often overlooked entirely and they get described as

$R(\alpha|\mu)$: due to fitting $\alpha$, ignoring $\beta$,

and

$R(\beta|\mu, \alpha)$:due to fitting $\beta$, adjusted for $\alpha$.

The omission of $\mu$ from descriptions such as these arises from a desire for verbal convenience. The omission is made with the convention that $\mu$ is not being ignored, even though it is not being mentioned. However, inclusion of $\mu$ in the descriptions is somewhat safer, for then there is no fear of its being overlooked. Furthermore, although in describing $R(\alpha|\mu)$, the phrase "ignoring $\beta$" is clear and appropriate, the phrase "adjusted for $\alpha$" in describing $R(\beta|\mu, \alpha)$ is not appealing because it may conjure up the idea of adjusting or amending the data in some manner. Since the

concept involved is clearly that of fitting $\beta$ over and above having fitting $\mu$ and $\alpha$, the description "$\beta$ after $\mu$ and $\alpha$" seems more appropriate. However, the relationship of such descriptions to those involving "ignoring $\alpha$" and "adjusted for $\beta$" should be borne in mind when encountering them in other texts. For example, just as $R(\alpha|\mu)$ of Tables 7.2b and 7.3b could be described as the sum of squares for fitting $\alpha$, adjusted for $\mu$ and ignoring $\beta$, so also could $R(\beta|\mu)$ of Tables 7.2c and 7.3c be called the sum of squares for fitting $\beta$, adjusted for $\mu$, and ignoring $\alpha$. However, the description of fitting $\beta$ after $\mu$ is preferred.

**(*vi*) *Interpretation of Results.*** From the preceding discussion, we see that $F(\alpha|\mu)$ and $F(\alpha|\mu, \beta)$ are not used for the same purpose. This is also true about $F(\beta|\mu)$ and $F(\beta|\mu, \alpha)$. Distinguishing between these two $F$'s is of paramount importance because it is a distinction that occurs repeatedly in fitting other models. Furthermore, the distinction does not exist with the familiar balanced data situation because then, as we shall see subsequently, $F(\alpha|\mu) = F(\alpha|\mu, \beta)$, and $F(\beta|\mu) = F(\beta|\mu, \alpha)$. The two $F$'s are not equal only for unbalanced data. They are always unequal for unbalanced data! These $F$-tests are not the same. The test based on the statistic $F(\alpha|\mu)$ is testing the effectiveness (in terms of explaining variation in $y$) of adding $\alpha$-effects to the model over and above $\mu$. The statistic $F(\alpha|\mu, \beta)$ tests the effectiveness of adding $\alpha$-effects to the model over and above having $\mu$ and $\beta$-effects in it. These tests are not the same, and neither of them should be described albeit loosely as "testing $\alpha$-effects." We must describe these tests more completely. The test associated with the statistic $F(\alpha|\mu)$ must be described the one "testing $\alpha$ after $\mu$." Likewise, the test associated with the statistic $F(\alpha|\mu, \beta)$ must be described as the one "testing $\alpha$ after $\mu$ and $\beta$." Similarly, $F(\beta|\mu)$ and $F(\beta|\mu, \alpha)$ are not the same. The statistic $F(\beta|\mu)$ tests "$\beta$ after $\mu$." The statistic $F(\beta|\mu, \alpha)$ tests "$\beta$ after $\mu$ and $\alpha$." Further distinction between $F$-statistics of this nature will become evident when we consider tests of linear hypotheses to which they relate.

In Table 7.2, all of the $F$-statistics are judged significant at the 5% level. As a result, we conclude that the $\alpha$-effects and the $\beta$-effects add materially to the exploratory power of the model. However, with other data, we may not be able to draw conclusions as easily. For example, suppose that in some set of data analogous to Table 7.2b, $F(\alpha|\mu)$ and $F(\beta|\mu, \alpha)$ were both significant but that, analogous to Table 7.2c, neither $F(\beta|\mu)$ nor $F(\alpha|\mu, \beta)$ were. Admittedly, this may happen with only very few sets of data. However, since computed $F$-statistics are functions of data, it is certainly *possible* for such an apparent inconsistency to occur. There then arises the problem of trying to draw conclusions from such a result. To do so is not always easy. As a result, the ensuing discussion of possible conclusions might not receive universal approval. The problems of interpretation that we shall discuss here receive scant mention in most texts. The reason is that they are not definitive, perhaps subject to personal judgment and certainly to knowledge of the data being analyzed. Furthermore, they are not amenable to exact mathematical treatment. Nevertheless, since they *are* problems of interpretation, they arise, in one way or another whenever data are analyzed. For this reason, it is worthwhile to reflect on what conclusions might be appropriate

in different situations. In attempting to do so, we are all too well aware of leaving ourselves wide open for criticism. However, at the very worst, exposition of the problems might be of some assistance.

The general problem we consider is what conclusions can be drawn from the various combinations of results that can arise as a result of the significance or non-significance of $F(\alpha|\mu)$, $F(\beta|\mu,\alpha)$, $F(\beta|\mu)$, and $F(\alpha|\mu,\beta)$ implicit in Tables 7.3b and 7.3c and illustrated in Tables 7.2b and 7.2c. First, these $F$-statistics should be considered only if $F(R_{\mathrm{m}}) = F(\alpha|\mu,\beta)$ of Table 7.3a is significant. This is so because it is only the significance of $F(R_{\mathrm{m}})$ which suggests that simultaneous fitting of $\alpha$ and $\beta$ has exploratory value for the variation in $y$. However, it does not necessarily mean that *both* $\alpha$ and $\beta$ are needed in the model. It is the investigation of this aspect of the model that arises from looking at $F(\alpha|\mu)$, $F(\beta|\mu,\alpha)$, $F(\beta|\mu)$, and $F(\alpha|\mu,\beta)$. Table 7.4 shows that there are 16 different situations to consider. There are four possible outcomes for $F(\alpha|\mu)$ and $F(\beta|\mu,\alpha)$. They are:

1. both $F$'s significant;
2. the statistic $F(\alpha|\mu)$ non-significant and $F(\beta|\mu,\alpha)$ significant;
3. the statistic $F(\alpha|\mu)$ significant and $F(\beta|\mu,\alpha)$ non-significant;
4. both $F$'s non-significant.

These are shown as row headings in Table 7.4. With each of these outcomes, four similar outcomes can occur for $F(\beta|\mu)$ and $F(\alpha|\mu,\beta)$. They are shown as column headings in Table 7.4. For each of the 16 resulting outcomes, the conclusion to be drawn is shown in the body of the table.

**TABLE 7.4 Suggested Conclusions According to the Significance (Sig) and Non-Significance (NS) of $F$-Statistics in Fitting a Model with Two Main Effects ($\alpha$'s and $\beta$'s)—See Table 7.3**

| | **Fitting $\beta$ and Then $\alpha$ After $\beta$** | | | | |
|---|---|---|---|---|---|
| Fitting $\alpha$ and then $\beta$ after $\alpha$ | $F(\beta\|\mu)$ :<br>$F(\alpha\|\beta,\mu)$ | Sig<br>Sig | NS<br>Sig | Sig<br>NS | NS<br>NS |
| **Effects to be included in the model** | | | | | |
| $F(\alpha\|\mu)$ :<br>$F(\beta\|\alpha,\mu)$ : | Sig<br>Sig | $\alpha$ and $\beta$ | $\alpha$ and $\beta$ | $\beta$ | Impossible |
| $F(\alpha\|\mu)$ :<br>$F(\beta\|\alpha,\mu)$ : | NS<br>Sig | $\alpha$ and $\beta$ | $\alpha$ and $\beta$ | $\beta$ | $\alpha$ and $\beta$ |
| $F(\alpha\|\mu)$ :<br>$F(\beta\|\alpha,\mu)$ : | Sig<br>NS | $\alpha$ | $\alpha$ | $\alpha$ and $\beta$ | $\alpha$ |
| $F(\alpha\|\mu)$ :<br>$F(\beta\|\alpha,\mu)$ : | NS<br>NS | Impossible | $\alpha$ and $\beta$ | $\beta$ | Neither $\alpha$ nor $\beta$ |

We now indulge in the verbal convenience of omitting $\mu$ from our discussion. Instead, we use phrases like "$\alpha$ being significant alone" for $F(\alpha|\mu)$ being significant and "$\alpha$ being significant after fitting $\beta$" for $F(\alpha|\mu, \beta)$ being significant. We do not use phrases like "$\alpha$ being significant" which does not distinguish between $F(\alpha|\mu)$ and $F(\alpha|\mu, \beta)$ being significant.

The first entry in Table 7.4 (row 1 column 1) corresponds to the case dealt within Table 7.2. There, both $\alpha$ and $\beta$ are significant when fitted alone or one after the other. Thus, the conclusion is to fit both. The entries in the first row and second column, or second column and first row are cases of both $\alpha$ and $\beta$ being significant when fitted after each other with one of them being significant when fitted alone, the other not. Again, the conclusion is to fit both. For the second diagonal entry (row 2 column 2), neither $\alpha$ nor $\beta$ is significant alone, but each is significant when fitted after the other. Hence, we fit both. Similarly, the entries in the third row and first column, or first column and third row are cases where one factor ($\beta$ in the third row and $\alpha$ in the third column) is significant only when fitted alone, but the other is significant either when fitted alone or after the first. Hence that other factor—$\alpha$ in the third row (first column) and $\beta$ in the third column (first row)—is the factor to fit. In the third row and second column, $\alpha$ and $\alpha$ after $\beta$ is significant but $\beta$ and $\beta$ after $\beta$ is not significant, so $\alpha$ is fitted. For the third column and second row, $\beta$ and $\beta$ after $\alpha$ is significant but $\beta$ and $\alpha$ after $\beta$ is not significant, so $\beta$ is fitted. For the third diagonal entry (row 3, column 3), both $\alpha$ and $\beta$ are significant alone but not after one another. Hence, we fit both $\alpha$ and $\beta$. If the model sum of squares is significant, it is impossible for both $\alpha$ and $\alpha$ after $\beta$, and $\beta$ and $\beta$ after $\alpha$ to not be significant. Hence, the fourth row first column, or first row fourth column is impossible to fit. For the fourth row second column, we have that only $\alpha$ after $\beta$ is significant, so we fit both $\alpha$ and $\beta$. Likewise, for the fourth column second row, we see that only $\beta$ after $\alpha$ is significant, so again, we fit both $\alpha$ and $\beta$. For the fourth row third column, only $\beta$ alone is significant, so $\beta$ is fitted. For the fourth column third row, only $\alpha$ alone is significant, so $\alpha$ is fitted. Finally, for the fourth diagonal entry (row 4 column 4), neither $\alpha$ nor $\beta$ is significant alone or after the other variable, so neither variable is fitted in the model.

In the case of the third diagonal entry, the decision to include both variables might be overridden; for example, if determining levels of the $\alpha$-factor was very costly, one might be prepared to use just the $\beta$-factor. Of course, this is a consideration that might arise with other entries in Table 7.4 too. The first two entries in the last row and column are difficult to visualize. Both pairs of entries are situations where fitting the factors in one sequence gives neither $F$-statistic significant, but fitting them in the other sequence gives the $F$-statistic for fitting the second factor significant. Intuitively, one feels that this kind of thing should happen somewhat infrequently. When it does, a reasonable conclusion seems to be to fit both factors as shown.[1]

In the widely used statistical package SAS, the sums of squares that result when variables are added sequentially to a model are called type I sum of squares. When one

[1] Grateful thanks go to N. S. Urquhart for lengthy discussions on this topic.

considers only the factors given all other factors in the model, the sums of squares are called type III sum of squares. The two sums of squares are different for unbalanced data but the same for balanced data. We illustrate this distinction in Example 2 below.

**Example 2 Type I and Type III Sum of Squares** Consider the SAS output below for the data of Table 7.1.

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **stove** | 4 | 1 2 3 4 |
| **pan** | 3 | 1 2 3 |
| **Number of Observations Read** | | 9 |
| **Number of Observations Used** | | 9 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: time**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 420.0000000 | 84.0000000 | 21.00 | 0.0153 |
| **Error** | 3 | 12.0000000 | 4.0000000 | | |
| **Corrected Total** | 8 | 432.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **time Mean** | |
| 0.972222 | 16.66667 | | 2.000000 | 12.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **stove** | 3 | 162.0000000 | 54.0000000 | 13.50 | 0.0301 |
| **pan** | 2 | 258.0000000 | 129.0000000 | 32.25 | 0.0094 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **stove** | 3 | 271.5000000 | 90.5000000 | 22.63 | 0.0146 |
| **pan** | 2 | 258.0000000 | 129.0000000 | 32.25 | 0.0094 |

In this instance, the $\alpha$-factor (brand of stove) was fitted first. The output corresponds to the results in in Table 7.2a and b. Under type I SS, we have $R(\alpha|\mu)$ and $R(\beta|\mu, \alpha)$, the associated $F$-statistics and $p$-values. Under type III SS, we have $R(\alpha|\mu, \beta)$ and $R(\beta|\mu, \alpha)$. As expected, $R(\alpha|\mu) \neq R(\alpha|\mu, \beta)$. Unlike Table 7.2, SAS omits the sum of squares due to the mean and computes the total sum of squares corrected for the mean.

We now look at the case where the $\beta$-factor (make of pan) is fitted first. This time, under type I SS, we have $R(\beta|\mu)$ and $R(\alpha|\mu, \beta)$. For the type III SS, we have $R(\beta|\mu, \alpha)$ and $R(\alpha|\mu, \beta)$, the associated $F$-statistics and $p$-values. Again, as expected, $R(\beta|\mu) \neq R(\beta|\mu, \alpha)$

Here is the SAS output fitting the $\beta$-factor first.

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **stove** | 4 | 1 2 3 4 |
| **pan** | 3 | 1 2 3 |
| **Number of Observations Read** | | 9 |
| **Number of Observations Used** | | 9 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: time**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 420.0000000 | 84.0000000 | 21.00 | 0.0153 |
| **Error** | 3 | 12.0000000 | 4.0000000 | | |
| **Corrected Total** | 8 | 432.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **time Mean** | |
| 0.972222 | 16.66667 | | 2.000000 | 12.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **pan** | 2 | 148.5000000 | 74.2500000 | 18.56 | 0.0204 |
| **stove** | 3 | 271.5000000 | 90.5000000 | 22.62 | 0.0146 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **pan** | 2 | 258.0000000 | 129.0000000 | 32.25 | 0.0094 |
| **stove** | 3 | 271.5000000 | 90.5000000 | 22.62 | 0.0146 |

The above output corresponds to that of Table 7.2a and c. It was generated by these commands.

```
Data boil;
Input stove pan time;
Cards;
1 1 18
1 2 12
..............
4 2 3
4 3 18
proc glm;
class stove pan;
model time =pan stove;
proc glm;
class stove pan;
model time =pan stove;
run;
```

The corresponding R output and program follows.

```
time<-c(18,12,24,NA,NA,9,3,NA,15,6,3,18)
> brand<-c("x","x","x","y","y","y","z","z","z","w","w","w")
> make<-c("a","b","c","a","b","c","a","b","c","a","b","c")
> resl<-lm(time~brand+make)
> anova(resl)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value  Pr(>F)
brand      3    162      54   13.50 0.03010 *
make       2    258     129   32.25 0.00937 **
Residuals  3     12       4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> resl1<-lm(time~make+brand)
> anova(resl1)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value  Pr(>F)
make       2  148.5   74.25  18.562 0.02044 *
brand      3  271.5   90.50  22.625 0.01459 *
Residuals  3   12.0    4.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

□

### f. Estimable Functions

The basic estimable function for the model (1) is

$$E(y_{ij}) = \mu + \alpha_i + \beta_j. \tag{42}$$

Its b.l.u.e. is

$$\mu + \hat{\alpha}_i + \beta_j = \mu^\circ + \alpha_i^\circ + \beta_j^\circ. \tag{43}$$

Note that although individual $\alpha$'s and $\beta$'s are not estimable, since linear functions of estimable functions are estimable, differences between pairs of $\alpha$'s and between pairs of $\beta$'s are estimable. Linear functions of these differences are also estimable. Thus, $\alpha_i - \alpha_h$ is estimable with b.l.u.e.

$$\widehat{\alpha_i - \alpha_h} = \alpha_i^\circ - \alpha_h^\circ, \tag{44a}$$

and $\beta_j - \beta_k$ is estimable with b.l.u.e.

$$\widehat{\beta_j - \beta_k} = \beta_j^\circ - \beta_k^\circ. \tag{44b}$$

The variances of these b.lu.e.'s are found from the general result for an estimable function $\mathbf{q}'\mathbf{b}$ that the variance of its b.l.u.e. is $v(\mathbf{q}'\mathbf{b}^\circ) = \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2$. Hence, if $g_{ii}$ and $g_{hh}$ are the diagonal elements of $\mathbf{G}$ corresponding to $\alpha_i$ and $\alpha_h$, respectively, and $g_{ih}$ is the element at the intersection of the row and column corresponding to $\alpha_i$ and $\alpha_h$, then

$$v(\widehat{\alpha_i - \alpha_h}) = v(\alpha_i^\circ - \alpha_h^\circ) = (g_{ii} + g_{hh} - 2g_{ih})\sigma^2. \tag{45}$$

A similar result holds for $v(\beta_j^\circ - \beta_k^\circ)$. Furthermore, any linear combination of the estimable functions in (44) is estimable with b.l.u.e., the same linear combination of the b.l.u.e.'s shown in (44). We can find variances of such b.l.u.e.'s in a manner similar to (45).

More generally, if $\mathbf{b} = \{b_s\}$ for $s = 1, 2, \ldots, a + b + 1$ and $\mathbf{G} = \{g_{s,t}\}$ for $s, t = 1, 2, \ldots, a + b + 1$ then, provided that $b_s - b_t$ is estimable (i.e., the difference of two $\alpha$'s or two $\beta$'s),

$$\widehat{b_s - b_t} = b_s^\circ - b_t^\circ, \text{ with } v(\widehat{b_s - b_t}) = (g_{ss} + g_{tt} - 2g_{st})\sigma^2. \tag{46}$$

**Example 3 The Variance of a Specific Estimable Function** In (11), we have $\alpha_1^\circ = 26$ and $\alpha_3^\circ = 14$. Thus, from (44),

$$\widehat{\alpha_1 - \alpha_3} = \alpha_1^\circ - \alpha_3^\circ = 26 - 14 = 12.$$

We earlier derived

$$\mathbf{G} = \frac{1}{12}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7 & 0 & 2 & 2 & -4 & -5 & 0 \\ 0 & 0 & 12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 8 & 2 & -4 & -2 & 0 \\ 0 & 3 & 0 & 2 & 7 & -4 & -5 & 0 \\ 0 & -4 & 0 & -4 & -4 & 8 & 4 & 0 \\ 0 & -5 & 0 & -2 & -5 & 4 & 11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{47}$$

Thus,

$$v(\widehat{\alpha_1 - \alpha_3}) = \frac{1}{12}[7 + 8 - 2(2))]\sigma^2 = \frac{11}{12}\sigma^2.$$

With $\sigma^2$ estimated as $\hat{\sigma}^2 = 4 = \text{MSE}$ in Table 7.2, the estimated variance is

$$v(\widehat{\alpha_1 - \alpha_3}) == \frac{11}{12}(4) = 2.6667. \qquad \square$$

### g. Tests of Hypotheses

As usual, the $F$-statistic for testing testable hypotheses $H$: $\mathbf{K}'\mathbf{b} = \mathbf{0}$ is

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{(\mathbf{K}'\mathbf{b}^\circ)'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ}{s\hat{\sigma}^2}$$

where,

$$Q = (\mathbf{K}'\mathbf{b}^\circ)'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ,$$

using (21) for $\mathbf{G}$, $s$ being the rank and number of rows of $\mathbf{K}'$.

In previous sections, we dealt at length with the meaning of the sums of squares in Table 7.2 and 7.3, interpreting them in terms of reductions in sums of squares due to fitting different models. We now consider their meaning in terms of testing hypotheses. In this context, there is no question of dealing with different models. We are testing hypotheses about the elements of the model (1). First, we show that $F(\beta|\alpha, \mu)$ of Table 7.2b is the $F$-statistic for testing the hypothesis that all the $\beta$'s are equal. If we state the hypothesis as $H$: $\beta_j - \beta_b = 0 \quad \text{for } j = 1, 2, \ldots, b-1$, the hypothesis can be written as

$$H\text{: } \mathbf{K}'\mathbf{b} = \mathbf{0} \quad \text{with } \mathbf{K}' = \begin{bmatrix} 0 & \mathbf{1}_{b-1} & \mathbf{0} & \mathbf{I}_{b-1} & -\mathbf{1}_{b-1} \end{bmatrix},$$

wherein, $\mathbf{K}'$ is partitioned conformably for the product $\mathbf{K}'\mathbf{G}$. Then, with $\mathbf{G}$ of (21)

$$\mathbf{K}'\mathbf{G} = \begin{bmatrix} 0 & -\mathbf{C}^{-1}\mathbf{M}' & \mathbf{C}^{-1} & \mathbf{0} \end{bmatrix}$$

and

$$\mathbf{K}'\mathbf{G}\mathbf{K} = \mathbf{C}^{-1}.$$

Furthermore,

$$\mathbf{K}'\mathbf{b}^\circ = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{y} = (-\mathbf{C}^{-1}\mathbf{M}'\mathbf{y}_a + \mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y}_\beta),$$

where $\mathbf{y}_a$ is as in (19), the vector of totals for the $a$ levels of the $\alpha$-factor and $\mathbf{y}_\beta = \{y_{.j}\}$ for $j = 1, \ldots, b-1$ is the vector of totals for the first $b-1$ levels of the

$\beta$-factor as in (22). Then, the numerator sum of squares of $F(H)$ is

$$\begin{aligned} Q &= (\mathbf{K}'\mathbf{b}^\circ)'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ \\ &= (-\mathbf{C}^{-1}\mathbf{M}'\mathbf{y}_a + \mathbf{C}^{-1}\mathbf{y}_\beta)'(\mathbf{C}^{-1})^{-1}(-\mathbf{C}^{-1}\mathbf{M}'\mathbf{y}_a + \mathbf{C}^{-1}\mathbf{y}_\beta) \\ &= (\mathbf{y}_\beta - \mathbf{N}'\mathbf{D}_a^{-1}\mathbf{y}_a)'\mathbf{C}^{-1}(\mathbf{y}_\beta - \mathbf{N}'\mathbf{D}_a^{-1}\mathbf{y}_a) \\ &= \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}, \text{ by the definition of } \mathbf{r} \text{ in (18)} \\ &= \beta^{\circ\prime}\mathbf{r}, \text{ by (16)} \\ &= R(\beta|\mu,\alpha) \text{ by (32).} \end{aligned}$$

**Example 4 Testing the Equality of the $\beta$'s** The hypothesis of equality of the $\beta$'s for the data in Table 7.1 can be written

$$\begin{aligned} H\colon \beta_1 - \beta_3 &= 0 \\ \beta_2 - \beta_3 &= 0. \end{aligned}$$

Using matrices it can be written as

$$\mathbf{K}'\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \mathbf{b} = \mathbf{0}.$$

With $\mathbf{b}^\circ$ of (11) and $\mathbf{G}$ of (47),

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} -10 \\ -14 \end{bmatrix}$$

and

$$\mathbf{K}'\mathbf{G} = \frac{1}{12}\begin{bmatrix} 0 & -4 & 0 & -4 & -4 & 8 & 4 & 0 \\ 0 & -5 & 0 & -2 & -5 & 4 & 11 & 0 \end{bmatrix}.$$

Thus,

$$\mathbf{K}'\mathbf{G}\mathbf{K} = \frac{1}{12}\begin{bmatrix} 8 & 4 \\ 4 & 11 \end{bmatrix} \quad \text{and} \quad (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \frac{1}{6}\begin{bmatrix} 11 & -4 \\ -4 & 8 \end{bmatrix}.$$

Hence, the numerator sum of squares of $F(H)$ is

$$\begin{aligned} Q = (\mathbf{K}'\mathbf{b}^\circ)'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ &= \begin{bmatrix} -10 & -14 \end{bmatrix} \frac{1}{6}\begin{bmatrix} 11 & -4 \\ -4 & 8 \end{bmatrix}\begin{bmatrix} -10 \\ -14 \end{bmatrix} \\ = 258 = R(\beta|\mu,\alpha) \text{ of Table 7.2b.} & \end{aligned}$$

□

The result of Example 4 can be obtained by stating the hypothesis in another, different but equivalent way. We illustrate this in Example 5.

**Example 5 Another Way to Test the Equality of the $\beta$'s** Another way to state the hypothesis that was tested in Example 4 is

$$\begin{aligned} H\colon \beta_1 - \beta_2 &= 0 \\ \beta_1 - \beta_3 &= 0 \end{aligned}$$

Now the matrix

$$\mathbf{K}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}.$$

Hence,

$$\mathbf{K}'\mathbf{b} = \begin{bmatrix} 4 \\ -10 \end{bmatrix}, \quad \mathbf{K}'\mathbf{G}\mathbf{K} = \frac{1}{12}\begin{bmatrix} 11 & 4 \\ 4 & 8 \end{bmatrix}, \quad \text{and} \quad (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \frac{1}{6}\begin{bmatrix} 8 & -4 \\ -4 & 11 \end{bmatrix}.$$

Then the numerator sum of squares for $F(H)$

$$Q = \begin{bmatrix} 4 & -10 \end{bmatrix}\frac{1}{6}\begin{bmatrix} 8 & -4 \\ 4 & 11 \end{bmatrix}\begin{bmatrix} 4 \\ -10 \end{bmatrix} = 258$$

as in Example 4. □

Thus $R(\beta|\mu, \alpha)$ is the numerator sum of squares for the $F$-statistic for testing $H$: all $\beta$'s equal. Similarly, $R(\alpha|\mu, \beta)$ is the numerator sum of squares for the $F$-statistic for testing $H$: all $\alpha$'s equal.

We can show by a similar argument that $R(\beta|\mu)$ is a numerator sum of squares for testing

$$H\colon \text{ equality of } \beta_j + \frac{1}{n_{\cdot j}}\sum_{j=1}^{a} n_{ij}\alpha_i \quad \text{for all} \quad j = 1, 2, \ldots, b. \tag{48a}$$

In Example 6 below, we demonstrate the test of this hypothesis for the data of Table 7.1.

**Example 6 Test of the Hypothesis in (48) for the Data of Table 7.1** The hypothesis can be conveniently stated as

$$\begin{aligned} H\colon \beta_1 + \tfrac{1}{3}(\alpha_1 + \alpha_3 + \alpha_4) - [\beta_3 + \tfrac{1}{4}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)] &= 0 \\ \beta_2 + \tfrac{1}{2}(\alpha_1 + \alpha_4) - [\beta_3 + \tfrac{1}{4}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)] &= 0. \end{aligned}$$

In matrix format we can write $H$ as

$$\mathbf{K}'\mathbf{b} = \begin{bmatrix} 0 & \frac{1}{12} & -\frac{1}{4} & \frac{1}{12} & \frac{1}{12} & 1 & 0 & -1 \\ 0 & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & 0 & 1 & -1 \end{bmatrix} \mathbf{b} = \mathbf{0}.$$

With $\mathbf{b}^\circ$ of (11) and $\mathbf{G}$ of (47) 23 have

$$\mathbf{K}'\mathbf{b}^\circ = \begin{bmatrix} -7.5 \\ -9 \end{bmatrix}, \quad \mathbf{K}'\mathbf{G}\mathbf{K} = \frac{1}{12}\begin{bmatrix} 7 & 3 \\ 3 & 9 \end{bmatrix} \quad \text{and} \quad (\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \frac{2}{9}\begin{bmatrix} 9 & -3 \\ -3 & 7 \end{bmatrix}.$$

Hence,

$$\begin{aligned} Q = (\mathbf{K}'\mathbf{b}^\circ)'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ &= \begin{bmatrix} -7.5 & -9 \end{bmatrix} \frac{2}{9}\begin{bmatrix} 9 & -3 \\ -3 & 7 \end{bmatrix}\begin{bmatrix} -7.5 \\ -9 \end{bmatrix} \\ &= 148.5 = R(\beta|\mu) \text{ of Table 7.2c.} \end{aligned}$$

□

In Example 6, we have demonstrated that $F(\beta|\mu) = R(\beta|\mu)/(b-1)\hat{\sigma}^2$. This is true in general. The $F$-statistic having $R(\alpha|\mu)$ of Table 7.2b as its numerator sum of squares, $F(\alpha|\mu)$, tests a hypothesis analogous to that in (48a), namely,

$$H\colon \alpha_i + \frac{1}{n_{i\cdot}}\sum_{j=1}^{b} n_{ij}\beta_j \text{ equal for all } i. \tag{48b}$$

The importance of these results is that $F(\alpha|\mu)$ is not a statistic for testing the equality of the $\alpha$'s. The statistic to be used for this test is $F(\alpha|\mu, \beta)$. The hypothesis that is tested by $F(\alpha|\mu)$ is the equality of the $\alpha$'s plus weighted averages of the $\beta$'s, the weights being the $n_{ij}$. Similarly, $F(\beta|\mu)$ does not test the equality of the $\beta$'s. It tests equality of the $\beta$'s plus weighted averages of the $\alpha$'s, as in (48a).

When using SAS, the relevant $p$- values for testing equality of the $\alpha$'s or the $\beta$'s are with the type III sum of squares. To test the hypothesis in (48a), use the type I sum of squares fitting $\beta$ first. Likewise, to test the hypothesis in (48b), again, use the type I sum of squares fitting $\alpha$ first.

### h. Models that Include Restrictions

Since $\mu + \alpha_i + \beta_j$ is estimable, so is $\mu + \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \beta_j$. Therefore, if the model includes the restriction that $\sum_{i=1}^{a}\alpha_i = 0$, then $\mu + \beta_j$ is estimable with b.l.u.e. $\mu^\circ + \frac{1}{a}\sum_{i=1}^{a}\alpha_i^\circ + \beta_j^\circ$, the same as the b.l.u.e. of $\mu + \frac{1}{a}\sum_{i=1}^{a}\alpha_i + \beta_j$ in the unrestricted model. Whether the restriction $\sum_{i=1}^{a}\alpha_i = 0$ is part of the model or not, the estimable functions and their b.l.u.e.'s given in (44) still apply. Thus, $\alpha_i^\circ - \alpha_h^\circ$ is still the b.l.u.e.

of $\alpha_i - \alpha_h$ and $\beta_j^\circ - \beta_k^\circ$ is the b.l.u.e. of $\beta_j - \beta_k$. Similar results hold if the model includes the restriction $\sum_{j=1}^{b} \beta_j = 0$.

The hypothesis of equality of $\beta_j + \frac{1}{n_{.j}} \sum_{i=1}^{a} n_{ij}\alpha_i$ for all $j = 1, 2, \ldots, b$ discussed in Section 1g might hint at the possibility of using a model that included the restriction

$$\sum_{i=1}^{a} n_{ij}\alpha_i = 0 \quad \text{for all} \quad j = 1, 2, \ldots, b. \tag{49}$$

Any value to this suggestion is lost whenever $b \geq a$. Then, equations (49) can be solved for the $\alpha$'s. In fact, $\alpha_i = 0$ for all $i$, regardless of the data. When $b < a$, equations (49) could be used as restrictions on the model. However, then only $a - b$ linear functions of the $\alpha$'s would be estimable from the data. Furthermore, since equations (49) are data-dependent, in that they are based on the $n_{ij}$, they suffer from the same deficiencies as do all such restrictions as explained at the end of Section 2h of Chapter 6.

### i. Balanced Data

The preceding discussion uses $n_{ij}$ as the number of observations in the $i$th row (level of the $\alpha$ factor) and the $j$th column level (level of the $\beta$ factor) with all $n_{ij} = 0$ or 1. Later, we will show that much of that discussion applies *in toto* to discussions where $n_{ij}$ can be any non-negative integers and hence to $n_{ij} = n$. However, here we consider the simplest case of balanced data where $n_{ij} = 1$ for all $i$ and $j$. These data are like that of Table 7.1, only without missing observations.

As one might expect, there is great simplification of the foregoing results when $n_{ij} = 1$ for all $i$ and $j$. Of course, the simplifications lead exactly to the familiar calculations in this case (e.g., see p. 72 of Kempthorne (1952), pp. 244–248 of Gruber (2014)). One may obtain a variety of solutions to the normal equations under these conditions. Using the procedure already given for unbalanced data which involves solving equations (14) and (15) (See Exercises 21 and 22), we obtain

$$\mu^\circ = 0, \quad \text{and} \quad \alpha_i^\circ = \bar{y}_{i.} - \bar{y}_{..} + \bar{y}_{.b} \quad \text{for all } i;$$

and

$$\beta_b^\circ = 0, \quad \text{and} \quad \beta_j^\circ = \bar{y}_{.j} - \bar{y}_{.b} \quad \text{for} \quad j = 1, 2, \ldots, b-1.$$

By use of the "usual constraints"

$$\sum_{i=1}^{a} \alpha_i^\circ = 0 = \sum_{j=1}^{b} \beta_j^\circ,$$

we obtain another set of solutions. They are

$$\mu^\circ = \bar{y}_{..},$$
$$\alpha_i^\circ = \bar{y}_{i.} - \bar{y}_{..} \quad \text{for all } i$$

and

$$\beta_j^\circ = \bar{y}_{.j} - \bar{y}_{..} \quad \text{for all } j.$$

In either case, the b.l.u.e.'s of differences between $\alpha$'s and $\beta$'s are

$$\widehat{\alpha_i - \alpha_h} = \bar{y}_{i.} - \bar{y}_{h.}, \quad \text{with } v(\widehat{\alpha_i - \alpha_h}) = \frac{2\sigma^2}{b}$$

and

$$\widehat{\beta_j - \beta_k} = \bar{y}_{.j} - \bar{y}_{.k}, \quad \text{with } v(\widehat{\beta_j - \beta_k}) = \frac{2\sigma^2}{a}.$$

Differences of this nature are always estimable. If the model includes restrictions $\sum_{i=1}^{a} \alpha_i = 0 = \sum_{j=1}^{b} \beta_j$ paralleling the usual constraints, then $\mu$, the $\alpha_i$ and $\beta_j$ are also estimable with $\hat{\mu} = \bar{y}_{..}$, $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$, and $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$.

The most noteworthy consequence of balanced data (all $n_{ij} = 1$) is that Tables 7.3b and 7.3c become identical for *all* data. This is the most important outcome of balanced data. It means that the distinction between $R(\alpha|\mu)$ and $R(\alpha|\mu, \beta)$ made in Tables 7.2 and 7.3 no longer occurs, because these two terms both simplify to be the same. Likewise, in these tables, there is no longer a distinction between $R(\beta|\mu)$ and $R(\beta|\mu, \alpha)$. They too simplify to be identically equal. The type I and type III sum of squares in SAS are equal for balanced data. Thus, when all $n_{ij} = 1$,

$$R(\alpha|\mu) = R(\alpha|\beta, \mu) = \sum_{i=1}^{a} \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{ab} = b \sum_{i=1}^{a} \bar{y}_{i.}^2 - ab\bar{y}_{..}^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{i.} - \bar{y}_{..})^2 \quad (50a)$$

and

$$R(\beta|\mu) = R(\beta|\mu, \alpha) = \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} = a \sum_{j=1}^{b} \bar{y}_{.j}^2 - ab\bar{y}_{..}^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2. \quad (50b)$$

The analysis of variance table becomes as shown in Table 7.5. When computing the sums of squares using a hand-held calculator, it is best to use the first of the three equivalent formulae in (50a) or (50b) to reduce round-off error. The sums of squares in (50a) and (50b) are well-known familiar expressions. Furthermore, they each have a simple form. They are easily calculated. They do not involve any matrix manipulations

**TABLE 7.5 Analysis of Variance for a Two-Way Classification with No Interaction with Balanced Data, all $n_{ij} = 1$. (Tables 7.3b and 7.3c Both Simplify to This Form When All $n_{ij} = 1$)**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = R(\mu) = ab\bar{y}_{..}^2$ |
| $\alpha$ after $\mu$ | $a - 1$ | $R(\alpha\|\mu) = R(\alpha\|\mu, \beta) = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{i.} - \bar{y}_{..})^2$ |
| $\beta$ after $\mu$ | $b - 1$ | $R(\beta\|\alpha, \mu) = R(\beta\|\mu) = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$ |
| Residual error | $(a - 1)(b - 1)$ | $\text{SSE} = \text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$ |
| Total | $ab$ | $\text{SST} = \text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij}^2$ |

like those previously described for unbalanced data (e.g., for $R(\beta|\mu, \alpha)$). In addition, because there is no longer any distinction between, for example, $R(\alpha|\mu)$ and $R(\alpha|\mu, \beta)$, there is no need to distinguish between fitting "$\alpha$ after $\mu$" and "$\alpha$ after $\mu$ and $\beta$." Our sole concern is with fitting "$\alpha$ after $\mu$" and similarly "$\beta$ after $\mu$." There is only one analysis of variance table, as shown in Table 7.5 where $R(\alpha|\mu)$ measures the efficacy of having the $\alpha$-effects in the model and $R(\beta|\mu)$ measures the efficacy of having the $\beta$ effects in it.

The convenience of a single analysis of variance table (Table 7.5) compared to having two analyses (Tables 7.3b and 7.3c) is obvious. For example, Table 7.4 is no longer pertinent. Unfortunately, however, this convenience that occurs with balanced data can easily result in a misunderstanding of the analysis of unbalanced data. Usually students encounter balanced data analysis first, such as that in Table 7.5. Explanation in terms of sums of squares of means $\bar{y}_{i.}$ (and $\bar{y}_{.j}$) about the general mean $\bar{y}_{..}$ has much intuitive appeal.However, unfortunately it does not carry over to unbalanced data. It provides, for example, no explanation as to why there are two analyses of variance for unbalanced data for a two-way classification. These two analyses have different meanings and are calculated differently (see Tables 7.3b and 7.3c). Furthermore, the calculations are quite different from those for balanced data. The manner of interpreting results is also different. In one analysis, we fit "$\alpha$ after $\mu$ and $\beta$." In the other, we fit "$\alpha$ after $\mu$." Small wonder that a student may experience disquiet when he views this state of affairs in the light of what has been arduously learned about balanced data. The changes to be made in the analysis and its interpretation appear so large in relation to the cause of it all—having unbalanced instead of balanced data—that the relationship of the analysis for unbalanced data to that for balanced data might, at least initially seem at all clear. The relationship is that balanced data are special cases of unbalanced data, and *not* vice versa.

**Example 7 A Two-Way Analysis of Variance for Balanced Data** The data below consist of the number of speeding-related fatalities, divided by the number of licensed

drivers in each of the listed states. These statistics were obtained from the Department of Transportation Highway Statistics.

| State/Speed limit (mph) | 55 | 50 | 45 | 40 | 35 | < 35 |
|---|---|---|---|---|---|---|
| California | 17.52 | 2.56 | 6.27 | 4.72 | 7.74 | 6.89 |
| Florida | 6.20 | 1.01 | 11.62 | 3.80 | 6.20 | 5.81 |
| Illinois | 28.1 | 0.372 | 2.73 | 5.84 | 8.57 | 10.93 |
| New York | 15.59 | 0.881 | 2.03 | 2.64 | 2.03 | 7.04 |
| Washington | 3.63 | 8.92 | 3.40 | 4.08 | 12.03 | 9.76 |

We have that,

$$y_{1.} = 45.7, y_{2.} = 34.64, y_{3.} = 56.543, y_{4.} = 30.211, y_{5.} = 41.52, y_{..} = 208.613,$$
$$y_{.1} = 71.04, y_{.2} = 13.443, y_{.3} = 26.05, y_{.4} = 21.08, y_{.5} = 36.57, y_{.6} = 40.43,$$
$$\text{SSM} = \frac{208.613^2}{30} = 1450.646$$

and the total sum of squares corrected for the mean is 981.43. Furthermore,

$$\text{SS(States)} = \frac{45.7^2 + 34.64^2 + 56.542^2 + 30.211^2 + 41.52^2}{6} - 1450.646 = 69.693$$

and

$$\text{SS(Speeds)} = \frac{71.04^2 + 13.443^2 + 26.05^2 + 21.08^2 + 36.57^2 + 40.43^2}{5} - 1450.646$$
$$= 413.817.$$

The residual error is SSE = 981.430 – 413.817 – 69.693 = 497.920.
The analysis of variance table is below.

| Source | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| States | 4 | 69.693 | 17.4231 | $0.70 < 3.38$ |
| Speed limits | 5 | 413.817 | 82.7634 | $3.32 > 2.71$ |
| Residual error | 20 | 497.920 | 24.8960 | |
| Total | 29 | 981.430 | | |

We conclude that there is not a significant difference in fatalities/per million drivers between states but there is amongst speed limits at the 5% level of significance.

Notice, in the R output below, that the order of fitting the two factors is unimportant for this balanced model.

Analysis of variance table

```
Response: fatal
          Df Sum Sq Mean Sq F value  Pr(>F)
speed      5 413.82  82.763  3.3244 0.02399 *
state      4  69.69  17.423  0.6998 0.60111
Residuals 20 497.92  24.896
-Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1--
Response: fatal
          Df Sum Sq Mean Sq F value  Pr(>F)
state      4  69.69  17.423  0.6998 0.60111
speed      5 413.82  82.763  3.3244 0.02399 *
Residuals 20 497.92  24.896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

□

**Example 8 Using the Rank Transformation for Speeding Fatalities Example** If we suspect the data are not normally distributed, we may use a rank transformation. The numbers in the table below in parenthesis are the relative ranks.

| State/Speed Limit (mph) | 55 | 50 | 45 | 40 | 35 | <35 |
|---|---|---|---|---|---|---|
| California | 17.52 (29) | 2.56 (6) | 6.27 (18) | 4.72 (13) | 7.74 (21) | 6.89 (19) |
| Florida | 6.20 (16.5) | 1.01 (3) | 11.62 (26) | 3.80 (11) | 6.20 (16.5) | 5.81 (14) |
| Illinois | 28.1 (30) | 0.372 (1) | 2.73 (8) | 5.84 (15) | 8.57 (22) | 10.93 (25) |
| New York | 15.59 (28) | 0.881 (2) | 2.03 (4.5) | 2.64 (7) | 2.03 (4.5) | 7.04 (20) |
| Washington | 3.63 (10) | 8.62 (23) | 3.40 (9) | 4.08 (12) | 12.03 (27) | 9.76 (24) |

The R output for analysis of variance on the ranks follows.

```
Analysis of Variance Table

Response: rank
          Df  Sum Sq Mean Sq F value  Pr(>F)
state      4  190.33  47.583  0.8104 0.53323
speedy     5  881.80 176.360  3.0035 0.03505 *
Residuals 20 1174.37  58.718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results and conclusions are similar to those obtained above using analysis of variance on the observations. □

**Example 9 Analysis of Variance for Data of Table 4.11 (Two-Way Without Interaction)**

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **refinery** | 3 | 1 2 3 |
| **source** | 4 | 1 2 3 4 |
| **Number of Observations Read** | | 25 |
| **Number of Observations Used** | | 25 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: percent**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 287.086338 | 57.417268 | 0.97 | 0.4631 |
| **Error** | 19 | 1129.473662 | 59.445982 | | |
| **Corrected Total** | 24 | 1416.560000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **percent Mean** | |
| 0.202664 | 20.70387 | | 7.710122 | 37.24000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **refinery** | 2 | 20.9627778 | 10.4813889 | 0.18 | 0.8397 |
| **source** | 3 | 266.1235602 | 88.7078534 | 1.49 | 0.2486 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **refinery** | 2 | 25.5346713 | 12.7673357 | 0.21 | 0.8087 |
| **source** | 3 | 266.1235602 | 88.7078534 | 1.49 | 0.2486 |

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **refinery** | 3 | 1 2 3 |
| **source** | 4 | 1 2 3 4 |
| **Number of Observations Read** | | 25 |
| **Number of Observations Used** | | 25 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: percent**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 5 | 287.086338 | 57.417268 | 0.97 | 0.4631 |
| **Error** | 19 | 1129.473662 | 59.445982 | | |
| **Corrected Total** | 24 | 1416.560000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **percent Mean** | |
| 0.202664 | 20.70387 | | 7.710122 | 37.24000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **source** | 3 | 261.5516667 | 87.1838889 | 1.47 | 0.2553 |
| **refinery** | 2 | 25.5346713 | 12.7673357 | 0.21 | 0.8087 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **source** | 3 | 266.1235602 | 88.7078534 | 1.49 | 0.2486 |
| **refinery** | 2 | 25.5346713 | 12.7673357 | 0.21 | 0.8087 |

```
Analysis of Variance Table

Response: percent
          Df  Sum Sq Mean Sq F value Pr(>F)
refinery   2   20.96  10.481  0.1763 0.8397
source     3  266.12  88.708  1.4922 0.2486
Residuals 19 1129.47  59.446
> anova(res2)
Analysis of Variance Table

Response: percent
          Df  Sum Sq Mean Sq F value Pr(>F)
source     3  261.55  87.184  1.4666 0.2553
refinery   2   25.53  12.767  0.2148 0.8087
Residuals 19 1129.47  59.446
```

□

As can be seen from the computer printouts, neither the source nor the refinery is statistically significant.

## 2. THE TWO-WAY CLASSIFICATION WITH INTERACTION

Suppose a plant-breeder carries out a series of experiments with three different fertilizer treatments on each of four varieties of grain. For each treatment-by-variety combination of plants, he plants several $4' \times 4'$ plots. At harvest time, he finds out that many of the plots have been lost due to having been wrongly plowed up. As a result, all he is left with is the data of Table 7.6. With four of the treatment-variety combination, there are no data at all. With the others, there are varying numbers of plots, ranging from one to four, with a total of 18 plots in all. Table 7.6 shows the yield of each plot, the total yields, the number of plots in each total, and the corresponding mean, for each treatment variety combination having data. Totals,

**TABLE 7.6 Weight[a] of Grain (Ounces) from $4' \times 4'$ Trial Plots**

| Treatment | Variety 1 | | | 2 | | | 3 | | | 4 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | | | | | | 12 | | | 7 | | | | | |
| | 13 | | | | | | | | | 11 | | | | | |
| | 9 | | | | | | | | | | | | | | |
| | 30 | (3) | 10[b] | | | | 12 | (1) | 12 | 18 | (2) | 9 | 60 | (6) | 10[b] |
| | $y_{11.}$ | $(n_{11})$ | $\bar{y}_{11.}$ | | | | $y_{13.}$ | $(n_{13})$ | $\bar{y}_{13.}$ | $y_{14.}$ | $(n_{14})$ | $\bar{y}_{14.}$ | $y_{1..}$ | $(n_{1.})$ | $\bar{y}_{1..}$ |
| 2 | 6 | | | 12 | | | | | | | | | | | |
| | 12 | | | 14 | | | | | | | | | | | |
| | 18 | (2) | 9 | 26 | (2) | 13 | | | | | | | 44 | (4) | 11 |
| | $y_{21.}$ | $(n_{21})$ | $\bar{y}_{21}$ | $y_{22.}$ | $(n_{22})$ | $\bar{y}_{22}$ | | | | | | | $y_{2..}$ | $(n_{2.})$ | $\bar{y}_{2..}$ |
| 3 | | | | 9 | | | 14 | | | 10 | | | | | |
| | | | | 7 | | | 16 | | | 14 | | | | | |
| | | | | | | | | | | 11 | | | | | |
| | | | | 16 | (2) | 8 | 30 | (2) | 15 | 48 | (4) | 12 | 94 | (8) | 11.75 |
| | | | | $y_{32.}$ | $(n_{32})$ | $\bar{y}_{32.}$ | $y_{33.}$ | $(n_{33})$ | $\bar{y}_{33.}$ | $y_{34.}$ | $(n_{34})$ | $\bar{y}_{34.}$ | $y_{3..}$ | $(n_{3.})$ | $\bar{y}_{3..}$ |
| Total | 48 | (5) | 9.6 | 42 | (4) | 10.5 | 42 | (3) | 14 | 66 | (6) | 11 | 198 | (18) | 11 |
| | $y_{.1.}$ | $(n_{.1})$ | $\bar{y}_{.1.}$ | $y_{.2.}$ | $(n_{.2})$ | $\bar{y}_{.2.}$ | $y_{.3.}$ | $n_{.3}$ | $\bar{y}_{.3.}$ | $y_{.4.}$ | $n_{.4}$ | $\bar{y}_{.4.}$ | $y_{...}$ | $n_{..}$ | $\bar{y}_{...}$ |

**(a) $n_{ij}$-values of Table 7.6**

| $i$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ | Total: $n_{i.}$ |
|---|---|---|---|---|---|
| 1 | 3 | 0 | 1 | 2 | 6 |
| 2 | 2 | 2 | 0 | 0 | 4 |
| 3 | 0 | 2 | 2 | 4 | 8 |
| Total: $n_{j}$ | 5 | 4 | 3 | 6 | $n = 18$ |

[a] The basic entries in the table are weights from individual plots.

[b] In each triplet of numbers, the first is the total weight, the second (in parentheses) is the number of plots in the total, and the third is the mean.

numbers of observations (plots), and means are shown for the three treatments, the four varieties, and for all 18 plots. The symbols for the entries in the table in terms of the model (see below) are also shown.

## a. Model

The equation of a suitable linear model for analyzing data of the kind in Table 7.6 is, as discussed in Chapter 4.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \tag{51}$$

The $y_{ijk}$ represents the $k$th observation in the $i$th treatment and $j$th variety. In (51), $\mu$ is a mean, $\alpha_i$ is the effect of the $i$th treatment, $\beta_j$ is the effect of the $j$th variety, $\gamma_{ij}$ is the interaction effect, and $e_{ijk}$ is the error term. In general, we have $\alpha_i$ as the effect due to the $i$th level of the $\alpha$-factor, $\beta_j$ is the effect due to the $j$th level of the $\beta$-factor, and $\gamma_{ij}$ is the interaction effect due to the $i$th level of the $\alpha$-factor and the $j$th level of the $\beta$-factor.

In the general case, there are $a$ levels of the $\alpha$-factor with $i = 1, \ldots, a$ and $b$ levels of the $\beta$-factor, with $j = 1, \ldots, b$. In the example of Table 7.6, $a = 3$ and $b = 4$.With balanced data, every one of the $ab$ cells in a table such as Table 7.6 would have $n$ observations. Furthermore, there would be $ab$ levels of the $\gamma$-factor (interaction factor) in the data. However, in unbalanced data, where some cells have no observations, as is the case in Table 7.6, there are only as many $\gamma$-levels in the data as there are non-empty cells. Let the number of such cells be $s$. In Table 7.6 $s = 8$. Then, if $n_{ij}$ is the number of observations in the $(i, j)$th cell (treatment $i$ and variety $j$), $s$ is the number of cells for which $n_{ij} \neq 0$ meaning that $n_{ij} > 0$ or $n_{ij} \geq 1$. For these cells,

$$y_{ij.} = \sum_{k=1}^{n_{ij}} y_{ijk}$$

is the total yield in the $(i, j)$th cell, and

$$\bar{y}_{ij.} = \frac{y_{ij.}}{n_{ij}}$$

is the corresponding mean. Similarly,

$$y_{i..} = \sum_{j=1}^{b} y_{ij.} \quad \text{and} \quad n_{i.} = \sum_{j=1}^{b} n_{ij}$$

are the total yield and the number of observations in the $i$th treatment. Corresponding values for the $j$th variety are

$$y_{.j.} = \sum_{i=1}^{a} y_{ij.} \quad \text{and} \quad n_{.j} = \sum_{i=1}^{a} n_{ij}.$$

The total yield for all plots is given by

$$y_{...} = \sum_{i=1}^{a} y_{i..} = \sum_{j=1}^{b} y_{.j.} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij.} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}.$$

The number of observations (plots) therein are

$$n_{..} = \sum_{i=1}^{a} n_{i.} = \sum_{j=1}^{b} n_{.j} = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}.$$

Examples of these symbols are shown in Table 7.6. The $n_{ij}$ notation used here is entirely parallel with that of the previous section, except that there $n_{ij} = 1$ or 0. Here $n_{ij} \geq 1$ or $n_{ij} = 0$.

The model equation $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ for the data of Table 7.6 is given in (52).

Equation (52) follows.

$$
\begin{array}{c}
\begin{bmatrix} 8 \\ 13 \\ 9 \\ 12 \\ 7 \\ 11 \\ 6 \\ 12 \\ 12 \\ 14 \\ 9 \\ 7 \\ 14 \\ 16 \\ 10 \\ 14 \\ 11 \\ 13 \end{bmatrix}
=
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{131} \\ y_{141} \\ y_{142} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{321} \\ y_{322} \\ y_{331} \\ y_{332} \\ y_{341} \\ y_{342} \\ y_{343} \\ y_{344} \end{bmatrix}
=
\begin{array}{c}
\begin{array}{cccccccccccccccc} \mu & \alpha_1 & \alpha_2 & \alpha_3 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \gamma_{11} & \gamma_{13} & \gamma_{14} & \gamma_{21} & \gamma_{22} & \gamma_{32} & \gamma_{33} & \gamma_{34} \end{array} \\
\begin{bmatrix}
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \gamma_{11} \\ \gamma_{13} \\ \gamma_{14} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{32} \\ \gamma_{33} \\ \gamma_{34} \end{bmatrix}
+
\begin{bmatrix} e_{111} \\ e_{112} \\ e_{113} \\ e_{131} \\ e_{141} \\ e_{142} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{321} \\ e_{322} \\ e_{331} \\ e_{332} \\ e_{341} \\ e_{342} \\ e_{343} \\ e_{344} \end{bmatrix}
\end{array}
\tag{52}
$$

**b. Normal Equations**

The normal equations $\mathbf{X'Xb}^\circ = \mathbf{X'y}$ corresponding to $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ are shown in (53).

$$
\begin{array}{c}
\begin{array}{cccccccccccccccc} \mu^\circ & \alpha_1^\circ & \alpha_2^\circ & \alpha_3^\circ & \beta_1^\circ & \beta_2^\circ & \beta_3^\circ & \beta_4^\circ & \gamma_{11}^\circ & \gamma_{13}^\circ & \gamma_{14}^\circ & \gamma_{21}^\circ & \gamma_{22}^\circ & \gamma_{32}^\circ & \gamma_{33}^\circ & \gamma_{34}^\circ \end{array} \\
\begin{bmatrix}
18 & 6 & 4 & 8 & 5 & 4 & 3 & 6 & 3 & 1 & 2 & 2 & 2 & 2 & 2 & 4 \\
6 & 6 & 0 & 0 & 3 & 0 & 1 & 2 & 3 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 4 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 0 \\
8 & 0 & 0 & 8 & 0 & 2 & 2 & 4 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 4 \\
5 & 3 & 2 & 0 & 5 & 0 & 0 & 0 & 3 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
4 & 0 & 2 & 2 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 \\
3 & 1 & 0 & 2 & 0 & 0 & 3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 \\
6 & 2 & 0 & 4 & 0 & 0 & 0 & 6 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 4 \\
3 & 3 & 0 & 0 & 3 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
2 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\
4 & 0 & 0 & 4 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4
\end{bmatrix}
\end{array}
\begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \beta_1^\circ \\ \beta_2^\circ \\ \beta_3^\circ \\ \beta_4^\circ \\ \gamma_{11}^\circ \\ \gamma_{13}^\circ \\ \gamma_{14}^\circ \\ \gamma_{21}^\circ \\ \gamma_{22}^\circ \\ \gamma_{32}^\circ \\ \gamma_{33}^\circ \\ \gamma_{34}^\circ \end{bmatrix}
=
\begin{bmatrix} y_{...} \\ y_{1..} \\ y_{2..} \\ y_{3..} \\ y_{.1.} \\ y_{.2.} \\ y_{.3.} \\ y_{.4.} \\ y_{11.} \\ y_{13.} \\ y_{14.} \\ y_{21.} \\ y_{22.} \\ y_{32.} \\ y_{33.} \\ y_{34.} \end{bmatrix}
=
\begin{bmatrix} 198 \\ 60 \\ 44 \\ 94 \\ 48 \\ 42 \\ 42 \\ 66 \\ 30 \\ 12 \\ 18 \\ 18 \\ 26 \\ 16 \\ 30 \\ 48 \end{bmatrix}
\end{array}
\tag{53}
$$

Properties of normal equations described in Section 4 of Chapter 6 are evident here. The first row of $\mathbf{X}'\mathbf{X}$ corresponding to the $\mu$-equation, has $n_{..}$ and the $n_{i.}$-, $n_{.j}$-, and $n_{ij}$-values as does the first column and the diagonal. The remaining elements of $\mathbf{X}'\mathbf{X}$, other than zeros are the $n_{ij}$-elements. The elements of $\mathbf{X}'\mathbf{y}$ on the right-hand side are all the totals $y_{...}, y_{i..}, y_{.j.}$, and $y_{ij.}$ shown in Table 7.6. As before, the partitioning of $\mathbf{X}'\mathbf{X}$ highlights its form.

### c. Solving the Normal Equations

The normal equations typified by (53) are easily solved. In (53), the number of equations is $p = 1 + a + b + s = 1 + 3 + 4 + 8 = 16$. The sum of the $\alpha$-equations (the three after the first) is identical to the $\mu$-equation. The sum of the $\beta$-equations enjoys the same property. Thus, there are two linear relationships among the rows of $\mathbf{X}'\mathbf{X}$. Furthermore, in the $\gamma$-equations, the sum of those pertaining to $\gamma_{i'j}$ summed over $j$ equals the $\alpha_{i'}$-equation. For example, the $\gamma_{11}$-, $\gamma_{13}$-, and $\gamma_{14}$-equations sum to the $\alpha_1$-equation. This holds true for all $i' = 1, 2, \ldots, a$, representing further linear relationships, $a$ of them, among rows of $\mathbf{X}'\mathbf{X}$. Similarly, in summing the $\gamma_{ij'}$-equations over $i$, the $\beta_{j'}$-equation is obtained for all $j' = 1, \ldots, b$. However, of the $b$ relationships represented here, only $b - 1$ are linearly independent of those already described. Thus, the total number of linearly independent relationships is $1 + 1 + a + b - 1 = 1 + a + b$. Hence the rank of $\mathbf{X}'\mathbf{X}$ is $r = 1 + a + b + s - (1 + a + b) = s$. Therefore, in terms of solving the normal equations by the procedure described in (4) of Chapter 6, we set $p - r = 1 + a + b + s - s = 1 + a + b$ elements of $\mathbf{b}^\circ$ equal to zero. The easiest elements for this purpose are $\mu^\circ$, all $\alpha_i^\circ$ ($a$ of them) and all $\beta_j^\circ$ ($b$ of them). Setting these equal to zero leaves, from (53), the $s = 8$ equations

$$\begin{aligned} 3\gamma_{11}^\circ &= 30, \quad 2\gamma_{22}^\circ = 26 \\ \gamma_{13}^\circ &= 12, \quad 2\gamma_{32}^\circ = 16 \\ 2\gamma_{14}^\circ &= 18, \quad 2\gamma_{33}^\circ = 30 \\ 2\gamma_{21}^\circ &= 18, \quad 4\gamma_{34}^\circ = 48. \end{aligned}$$

In general, the reduced equations are

$$n_{ij}\gamma_{ij}^\circ = y_{ij.}$$

with solution

$$\gamma_{ij}^\circ = \bar{y}_{ij.}, \tag{54}$$

for all the $(i, j)$ cells for which $n_{ij} \neq 0$, all $s$ of them. Then, the solution for $\mathbf{b}^\circ$ has the first $1 + a + b$ elements zero and the rest of the elements given by (54). Thus, we have the solution expressed as

$$\mathbf{b}^{\circ\prime} = \left[ \mathbf{0}_{1\times(1+a+b)} \; \overline{\mathbf{y}'} \right] \tag{55}$$

where $(\overline{\mathbf{y}'})_{1\times s} =$ a vector of all $\bar{y}_{ij.}$'s for which $n_{ij} \neq 0$.

In our example, writing out all of the components of the vector (55),

$$\begin{aligned}\mathbf{b}^{\circ\prime} &= \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{y}_{11.} & \bar{y}_{13.} & \bar{y}_{14.} & \bar{y}_{21.} & \bar{y}_{22.} & \bar{y}_{32.} & \bar{y}_{33.} & \bar{y}_{34.}\end{bmatrix} \\ &= \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 12 & 9 & 9 & 13 & 8 & 15 & 12\end{bmatrix} \end{aligned} \tag{56}$$

from Table 7.6.

The simplicity of this solution means that it is virtually unnecessary to derive the generalized inverse of $\mathbf{X}'\mathbf{X}$ that corresponds to $\mathbf{b}^{\circ}$. From (55) and the normal equations (53),

$$\mathbf{G} = \begin{bmatrix} \mathbf{0}_{(1+a+b)\times(1+a+b)} & \mathbf{0}_{(1+a+b)\times s} \\ \mathbf{0}_{s\times(1+a+b)} & \mathbf{D}\{1/n_{ij}\} \end{bmatrix} \tag{57}$$

where $\mathbf{D}\{1/n_{ij}\}$ is a diagonal matrix of order $s$ of the values $1/n_{ij}$ for the non-zero $n_{ij.}$

### d. Analysis of Variance

**(*i*) *Basic Calculations.*** The analysis of variance for the two-way classification model with interaction is similar to that for the two-way classification without interaction discussed in Section 1. Indeed, the analysis of variance tables are just like those of Tables 7.2 and 7.3, except for the inclusion of an interaction line corresponding to the sum of squares $R(\gamma|\mu,\alpha,\beta)$. Calculation of $R(\mu), R(\mu,\alpha), R(\mu,\beta),$ and $R(\mu,\alpha,\beta)$ is the same except for using $y_{...}, y_{i..}, y_{.j.},$ and $y_{ij.}$ respectively in place of $y_{..}, y_{i.}, y_{.j},$ and $y_{ij}$ used in the no-interaction model. Thus, similar to (24), (25), and (36),

$$R(\mu) = n_{..}\bar{y}_{...}^2 = \frac{y_{...}^2}{n_{..}}, \tag{58}$$

$$R(\mu,\alpha) = \sum_{i=1}^{a} n_{i.}\bar{y}_{i..}^2 = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_{.i}}, \tag{59}$$

and

$$R(\mu,\beta) = \sum_{j=1}^{b} n_{.j}\bar{y}_{.j.}^2 = \sum_{j=1}^{b} \frac{y_{.j.}^2}{n_{.j}}. \tag{60}$$

The model (51) involves the terms $\mu, \alpha_i, \beta_j,$ and$\gamma_{ij}$. The sum of squares for fitting it is therefore denoted by $R(\mu, \alpha, \beta, \gamma)$. Its value, as usual, is $\mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y}$. With $\mathbf{X}'\mathbf{y}$ of (53) and $\mathbf{b}^{\circ\prime}$ of (55), this gives

$$\begin{aligned} R(\mu, \alpha, \beta, \gamma) &= \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{y} \\ &= \overline{\mathbf{y}'}(\text{column vector of} y_{ij} \text{ total}) \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b} \overline{y}_{ij.} y_{ij.} \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\overline{y}_{ij.}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij.}^2}{n_{ij}}. \end{aligned} \tag{61}$$

In the last expression of (61), the terms $y_{ij}^2/n_{ij}$ are defined only for non-zero values of $n_{ij}$ in the data. The other term that we need for the analysis is $R(\mu, \alpha, \beta)$. This is the sum of squares due to fitting the reduced model

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}. \tag{62}$$

This is derived exactly as in equation (26). Thus,

$$R(\mu, \alpha, \beta) = \sum_{i=1}^{a} n_{i.}\overline{y}_{i..}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} \tag{63}$$

where,

$$\mathbf{C} = \{c_{jj'}\} \quad \text{for} \quad j, j' = 1, 2, \ldots, b-1 \tag{64}$$

with $c_{jj} = n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}}$, $c_{jj'} = -\sum_{i=1}^{n} \frac{n_{ij}n_{ij'}}{n_{i.}}$ for $j \neq j'$, and

$$\mathbf{r} = \{r_j\} = \left\{ y_{.j.} - \sum_{i=1}^{a} n_{ij}\overline{y}_{i..} \right\} \quad \text{for} \quad j = 1, 2, \ldots, b-1. \tag{65}$$

These are the same calculations as in (16)–(18) using $\overline{y}_{i..}$ and $\overline{y}_{.j.}$ in place of $\overline{y}_{i.}$ and $\overline{y}_{.j}$.

**Example 10 Sums of Squares for Data in Table 7.6** Calculation of (58)–(61) for the data of Table 7.6 is as follows

$$\begin{aligned} R(\mu) &= \frac{198^2}{18} = 2178, \\ R(\mu,\alpha) &= \frac{60^2}{6} + \frac{44^2}{4} + \frac{94^2}{8} = 2188.5, \\ R(\mu,\beta) &= \frac{48^2}{5} + \frac{42^2}{4} + \frac{42^2}{3} + \frac{66^2}{6} = 2215.8, \\ R(\mu,\alpha,\beta,\gamma) &= \frac{30^2}{3} + \frac{12^2}{1} + \cdots + \frac{30^2}{2} + \frac{48^2}{4} = 2260. \end{aligned} \tag{66}$$

As usual, the total sum of squares is

$$\sum y^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}^2 = 8^2 + 13^2 + \cdots + 11^2 + 13^2 = 2316. \tag{67}$$

To facilitate calculation of $R(\mu,\alpha,\beta)$, we use the table of $n_{ij}$'s shown in Table 7.6a. Using the values given there, $\mathbf{C}$ of (64) is

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} 5 - \left(\frac{3^2}{6} + \frac{2^2}{4}\right) & -\frac{2(2)}{4} & -\frac{1(3)}{6} \\ -\frac{2(2)}{4} & 4 - \left(\frac{2^2}{4} + \frac{2^2}{8}\right) & -\frac{2(2)}{8} \\ -\frac{1(3)}{6} & -\frac{2(2)}{8} & 3 - \left(\frac{1^2}{6} + \frac{2^2}{8}\right) \end{bmatrix} \\ &= \frac{1}{6}\begin{bmatrix} 15 & -6 & -3 \\ -6 & 15 & -3 \\ -3 & -3 & 14 \end{bmatrix}. \end{aligned}$$

Its inverse is

$$\mathbf{C}^{-1} = \frac{1}{126}\begin{bmatrix} 67 & 31 & 21 \\ 31 & 67 & 21 \\ 21 & 21 & 63 \end{bmatrix}.$$

From Table 7.6 and (65),

$$\mathbf{r} = \begin{bmatrix} 48 - 3(10) - 2(11) \\ 42 - 2(11) - 2(11.75) \\ 42 - 1(10) - 2(11.75) \end{bmatrix} = \begin{bmatrix} -4 \\ -3.5 \\ 8.5 \end{bmatrix}.$$

□

**TABLE 7.7 Analyses of Variance for Two-Way Classification with Interaction (Data of Table 7.6)**

| Source of Variation | d.f. | Sum of Squares | Mean Square | $F$-Statistic |
|---|---|---|---|---|
| **(a) For fitting $\mu$, and $\alpha, \beta, \gamma \mid \mu$** | | | | |
| Mean: $\mu$ | $1 = 1$ | $R(\mu) = 2,178$ | 2178 | $F(M) = 388.9$ |
| $\alpha, \beta,$ and $\gamma$ after $\mu$: $\alpha,\beta,\gamma \mid \mu$ | $s - 1 = 7$ | $R(\alpha, \beta, \gamma \mid \mu) = 82$ | 11.71 | $F(R_{\mathrm{m}}) = 2.1$ |
| Residual error | $N - s = 10$ | SSE = 56 | 5.60 | |
| Total | $N = 18$ | SST = 2316 | | |
| **(b) For fitting $\mu$, then $\alpha$, then $\beta$ and then $\gamma$: that is, for fitting $\mu, (\alpha \mid \mu), (\beta \mid \mu, \alpha)$ and $(\gamma \mid \mu, \alpha, \beta)$.** | | | | |
| Mean: $\mu$ | $1 = 1$ | $R(\mu) = 2,178$ | 2178 | $F(M) = 388.9$ |
| $\alpha$ after $\mu$: $\alpha \mid \mu$ | $a - 1 = 2$ | $R(\alpha \mid \mu) = 10.5$ | 5.25 | $F(\alpha \mid \mu) = 0.9$ |
| $\beta$ after $\mu$ and $\alpha$: $\beta \mid \mu, \alpha$ | $b - 1 = 3$ | $R(\beta \mid \alpha, \mu) = 36.79$ | 12.26 | $F(\beta \mid \alpha, \mu) = 2.2$ |
| $\gamma$ after $\mu,\alpha,$ and $\beta$:$\gamma \mid \mu, \alpha,$ and $\beta$ | $s - a - b + 1 = 2$ | $R(\gamma \mid \mu, \alpha, \beta) = 34.71$ | 17.36 | $F(\gamma \mid \alpha, \mu, \beta) = 3.1$ |
| Residual error | $N - \mathrm{s} = 10$ | SSE = 56 | 5.60 | |
| Total | $N = 18$ | SST = 2316 | | |
| **(c) For fitting $\mu$, then $\beta$, then $\alpha$, and then $\gamma$: that is, for fitting $\mu, (\beta \mid \mu), (\alpha \mid \mu, \beta)$ and $(\gamma \mid \mu, \alpha, \beta)$.** | | | | |
| Mean: $\mu$ | $1 = 1$ | $R(\mu) = 2,178$ | 2178 | $F(M) = 388.9$ |
| $\beta$ after $\mu$: $\beta \mid \mu$ | $b - 1 = 3$ | $R(\beta \mid \mu) = 37.8$ | 12.60 | $F(\beta \mid \mu) = 2.3$ |
| $\alpha$ after $\mu$ and $\beta$: $\alpha \mid \mu, \beta$ | $a - 1 = 2$ | $R(\alpha \mid \mu, \beta) = 9.49$ | 4.75 | $F(\alpha \mid \mu, \beta) = 0.8$ |
| $\gamma$ after $\mu, \beta,$ and $\alpha$: $\gamma \mid \mu, \alpha,$ and $\beta$ | $s - a - b + 1 = 2$ | $R(\gamma \mid \mu, \alpha, \beta) = 34.71$ | 17.36 | $F(\gamma \mid \alpha, \mu, \beta) = 3.1$ |
| Residual error | $N - \mathrm{s} = 10$ | SSE = 56 | 5.6 | |
| Total | $N = 18$ | SST = 2316 | | |

Therefore, (63), using $R(\mu, \alpha)$ from (66), gives

$$R(\mu, \alpha, \beta) = 2188.5 + \begin{bmatrix} -4 & -3.5 & -8.5 \end{bmatrix} \frac{1}{126} \begin{bmatrix} 67 & 31 & 21 \\ 31 & 67 & 21 \\ 21 & 21 & 63 \end{bmatrix} \begin{bmatrix} -4 \\ -3.5 \\ -8.5 \end{bmatrix} = 2225.29 \tag{68}$$

If, quite generally, one wishes to fit the model (62) from the beginning to data of the type illustrated in Table 7.6, the procedure just outlined yields the sum of squares for doing so, namely $R(\mu, \alpha, \beta)$. Thus, the procedure as described in Section 1 for calculating $R(\mu, \alpha, \beta)$ for the no-interaction model with $n_{ij} = 0$ or 1, is also the basis for calculating $R(\mu, \alpha, \beta)$ whenever the data are unbalanced. This is the case both when $R(\mu, \alpha, \beta)$ is needed as part of the analysis of variance for the interaction model and when the prime interest lies in $R(\mu, \alpha, \beta)$ itself as the reduction in the sum of squares due to fitting the no-interaction model (62).

**(*ii*) *Fitting Different Models.*** Analyses of variance derived from the sums of squares in (66), (67), and (68) are shown in Table 7.7. Their form is similar to that of Table 7.2. Table 7.7a shows the partitioning of the sum of squares $R(\mu, \alpha, \beta, \gamma)$ into two parts. They are $R(\mu)$ and $R(\alpha, \beta, \gamma|\mu)$ for fitting the $\alpha$-, $\beta$-, and $\gamma$-factors after the mean. For the data of Table 7.6, $R(\mu)$ is as shown in (66). This yields

$$R(\alpha, \beta, \gamma|\mu) = R(\mu, \alpha, \beta, \gamma) - R(\mu) = 2260 - 2178 = 82.$$

In the usual manner, the residual sum of squares is

$$\text{SSE} = \sum y^2 - R(\mu, \alpha, \beta, \gamma) = 2316 - 2260 = 56.$$

These are the terms shown in Table 7.7a. The corresponding $F$-statistics are also shown in Table 7.7a. We have that $F(M) = 388.9$ is significant because it exceeds 4.96, the 5% value of the $F_{1,10}$-distribution. Hence we reject the hypothesis $H$: $E(\overline{y}) = 0$. On the other hand, $F(R_{\text{m}}) = 2.1$ is less than the value of the $F_{7,10}$- distribution, namely, 3.14. As a result, we conclude that the $\alpha$-, $\beta$-, and $\gamma$-factors in the model are not effective in explaining the variation in the $y$'s over and above that explained by fitting a mean.

The data in Table 7.6 are hypothetical. In the analysis of variance in Table 7.7a, $F(R_{\text{m}}) = 2.1$ is not significant. Therefore, calculation of the analyses of variance shown in Tables 7.7b and 7.7c is therefore, not necessary. Nevertheless, it is instructive to examine the format of the analyses to see how similar they are to Tables 7.2b and 7.2c. Were $F(R_{\text{m}})$ of Table 7.7b significant, we would be led to examine whether it was the $\alpha$-factor, $\beta$-factor, $\gamma$-factor, or some combination thereof that was contributing to this significance. After fitting $\mu$, we could do this in one of two ways. We could either fit $\mu$ and $\alpha$, and then $\mu, \alpha,$ and $\beta$, or fit $\mu$ and $\beta$, and then fit $\mu, \alpha,$ and $\beta$. Either way, $\gamma$ would be fitted after having fitted $\mu, \alpha,$ and $\beta$. The choice lies in the first fitting after $\mu$, either $\alpha$ or $\beta$. This is exactly the situation discussed when describing Tables 7.2. Therefore, Tables 7.7b and 7.7c are similar in format to Tables 7.2b and 7.2c. Table 7.7b shows the partitioning of $R(\mu, \alpha, \beta, \gamma)$ for fitting $\mu$, then $\alpha$, then $\beta$ and then $\gamma$, with lines in the analysis of variance for $\mu$, $\alpha$ after $\mu$, $\beta$ after $\mu$ and $\alpha$, and finally $\gamma$ after $\mu, \alpha,$ and $\beta$. The only difference between Table 7.2 and Table 7.7b is that Table 7.7b contains the sum of squares $R(\gamma|\mu, \alpha, \beta)$. Of course, this corresponds to the $\gamma$-factor which is additional in the interaction model to the $\alpha$-and $\beta$-factors that are present in both the interaction and the no-interaction models. Using (68) and (69), we see that the sums of squares of Table 7.7b are

$$\begin{aligned} R(\mu) &= R(\mu) = 2178 = 2178, \\ R(\alpha|\mu) &= R(\mu, \alpha) - R(\mu) = 2188.5 - 2178 = 10.5 \\ R(\beta|\mu, \alpha) &= R(\mu, \alpha, \beta) - R(\mu, \alpha) = 2225.29 - 2188.5 = 36.79 \end{aligned}$$

and

$$R(\gamma|\mu, \alpha, \beta) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta) = 2225.29 - 2188.5 = 34.79.$$

With a slight error due to rounding off,[2] these sums do indeed add to 2260. Thus, they are a partitioning of this sum of squares. These results are shown in Table 7.7b. Naturally, $R(\mu) = 2178$ and $\text{SSE} = \sum y^2 - R(\mu, \alpha, \beta, \gamma) = 56$ are the same throughout Table 7.7. Moreover, the middle three entries of Table 7.7b add up to $R(\alpha, \beta, \gamma|\mu) = 82$, the middle entry of Table 7.7a, as do the middle entries of Table 7.7c. In this way, Tables 7.7b and 7.7c are partitionings not only of $R(\mu, \alpha, \beta, \gamma)$ but also of $R(\alpha, \beta, \gamma|\mu)$, the sum of the squares due to fitting the model over and above the mean.

The analogy between Tables 7.7b and Table 7.2b is repeated in Tables 7.7c and 7.2c. Thus, Table 7.7c has lines in the analysis of variance for $\mu$, $(\beta|\mu)$, $(\alpha|\mu, \beta)$, and $(\gamma|\mu, \alpha, \beta)$. The only difference from Table 7.7b is that $R(\alpha|\mu)$ and $R(\beta|\mu, \alpha)$ in Table 7.7b are replaced by $R(\beta|\mu)$ and $R(\alpha|\mu, \beta)$ in Table 7.7c. Observe that

$$R(\beta|\mu) = R(\mu, \beta) - R(\mu) = 2215.8 - 2178 = 37.8$$

and

$$R(\alpha|\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta) = 2225.29 - 2215.8 = 9.49.$$

Of course, the sum of these, except for a slight round-off error, is the same as the sum of $R(\alpha|\mu)$ and $R(\beta|\mu, \alpha)$ in Table 7.7b. Observe that

$$R(\beta|\mu) + R(\alpha|\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta) = R(\alpha|\mu) + R(\beta|\mu, \alpha).$$

Using values from Table 7.7 we have $37.8 + 9.49 = 2225.29 - 2178 = 10.5 + 36.79 = 47.29$. The sum in the above equation both for symbols and substituted numbers is $R(\mu, \alpha, \beta) - R(\mu) = R(\alpha, \beta|\mu)$.We now give an R output and program for computation of Tables 7.7b and 7.7c.

```
weight<-c(8,13,9,12,7,11,6,12,12,14,9,7,14,16,10,14,11,13)
> treatment<-c("ta","ta","ta","ta","ta","ta","tb","tb","tb","tb",
"tc","tc","tc","tc","tc","tc","tc","tc")
> variety<-c("va","va","va","vc","vd","vd","va","va","vb","vb","vb",
"vb","vc","vc","vd","vd","vd","vd")
> result1<-lm(weight~treatment*variety)
> result2<-lm(weight~variety*treatment)
> anova(result1)
Analysis of Variance Table

Response: weight
                  Df Sum Sq Mean Sq F value  Pr(>F)
treatment          2 10.500   5.250  0.9375 0.42348
variety            3 36.786  12.262  2.1896 0.15232
treatment:variety  2 34.714  17.357  3.0995 0.08965.
Residuals         10 56.000   5.600
---
```

[2] In the first edition, the first author used fractions. The second author chose to use decimals in revising the manuscript.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(result2)
Analysis of Variance Table

Response: weight
                  Df Sum Sq Mean Sq F value  Pr(>F)
variety            3 37.800 12.6000  2.2500 0.14507
treatment          2  9.486  4.7429  0.8469 0.45731
variety:treatment  2 34.714 17.3571  3.0995 0.08965.
Residuals         10 56.000  5.6000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The SAS output and program follows.

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **treatment** | 3 | 1 2 3 |
| **variety** | 4 | 1 2 3 4 |
| **Number of Observations Read** | | 18 |
| **Number of Observations Used** | | 18 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: weight**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 7 | 82.0000000 | 11.7142857 | 2.09 | 0.1400 |
| **Error** | 10 | 56.0000000 | 5.6000000 | | |
| **Corrected Total** | 17 | 138.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **weight Mean** | |
| 0.594203 | 21.51302 | | 2.366432 | 11.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **treatment** | 2 | 10.50000000 | 5.25000000 | 0.94 | 0.4235 |
| **variety** | 3 | 36.78571429 | 12.26190476 | 2.19 | 0.1523 |
| **treatment*variety** | 2 | 34.71428571 | 17.35714286 | 3.10 | 0.0897 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **treatment** | 2 | 12.47058824 | 6.23529412 | 1.11 | 0.3659 |
| **variety** | 3 | 34.87213740 | 11.62404580 | 2.08 | 0.1672 |
| **treatment*variety** | 2 | 34.71428571 | 17.35714286 | 3.10 | 0.0897 |

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **treatment** | 3 | 1 2 3 |
| **variety** | 4 | 1 2 3 4 |
| **Number of Observations Read** | | **18** |
| **Number of Observations Used** | | **18** |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: weight**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 7 | 82.0000000 | 11.7142857 | 2.09 | 0.1400 |
| **Error** | 10 | 56.0000000 | 5.6000000 | | |
| **Corrected Total** | 17 | 138.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **weight Mean** | |
| 0.594203 | 21.51302 | | 2.366432 | 11.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **variety** | 3 | 37.80000000 | 12.60000000 | 2.25 | 0.1451 |
| **treatment** | 2 | 9.48571429 | 4.74285714 | 0.85 | 0.4573 |
| **treatment*variety** | 2 | 34.71428571 | 17.35714286 | 3.10 | 0.0897 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **variety** | 3 | 34.87213740 | 11.62404580 | 2.08 | 0.1672 |
| **treatment** | 2 | 12.47058824 | 6.23529412 | 1.11 | 0.3659 |
| **treatment*variety** | 2 | 34.71428571 | 17.35714286 | 3.10 | 0.0897 |

```
The code to generate this output is
data grain;
input treatment variety weight;
cards;
1 1 8
1 1 13
1 1 9
.............
3 4 11
3 4 13
proc glm;
class treatment variety;
model weight =treatment variety treatment*variety;
proc glm;
class treatment variety;
model weight = variety treatment variety*treatment;
run;
```

An R output and program for Table 4.11 where process is taken as replication and interaction between refinery and source is included is given below.

```
> percentage<-c(31,33,44,36,38,26,37,59,42,42,34,42,28,39,36,32,38,
42,36,22,42,46,26,37,43)
> refinery<-c("g","g","g","g","g","g","g","g","g","n","n","n","n",
"n","n","n","n","s","s","s","s","s","s","s","s")
> source<-c("t","t","t","t","o","m","t","t","o","m","i","i","i","t",
"o","m","m","t","o","i","o","o","m","i","i")
> result1<-lm(percentage~refinery*source)
> result2<-lm(percentage~source*refinery)
> anova(result1)
Analysis of Variance Table
Response: percentage
               Df Sum Sq Mean Sq F value Pr(>F)
refinery        2  20.96  10.481  0.1507 0.8615
source          3 266.12  88.708  1.2751 0.3212
refinery:source 5 155.47  31.095  0.4469 0.8086
Residuals      14 974.00  69.571
> anova(result2)
Analysis of Variance Table

Response: percentage
               Df Sum Sq Mean Sq F value Pr(>F)
source          3 261.55  87.184  1.2532 0.3282
refinery        2  25.53  12.767  0.1835 0.8343
source:refinery 5 155.47  31.095  0.4469 0.8086
Residuals      14 974.00  69.571
```

Observe that source, refinery, and interaction are not significant. A SAS output is given below.

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **refinery** | 3 | 1 2 3 |
| **source** | 4 | 1 2 3 4 |
| **Number of Observations Read** | | 25 |
| **Number of Observations Used** | | 25 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: percent**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 10 | 442.560000 | 44.256000 | 0.64 | 0.7616 |
| **Error** | 14 | 974.000000 | 69.571429 | | |
| **Corrected Total** | 24 | 1416.560000 | | | |

| R-Square | Coeff Var | Root MSE | percent Mean |
|---|---|---|---|
| 0.312419 | 22.39782 | 8.340949 | 37.24000 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| refinery | 2 | 20.9627778 | 10.4813889 | 0.15 | 0.8615 |
| source | 3 | 266.1235602 | 88.7078534 | 1.28 | 0.3212 |
| refinery*source | 5 | 155.4736620 | 31.0947324 | 0.45 | 0.8086 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| refinery | 2 | 10.7659314 | 5.3829657 | 0.08 | 0.9259 |
| source | 3 | 282.6326238 | 94.2108746 | 1.35 | 0.2972 |
| refinery*source | 5 | 155.4736620 | 31.0947324 | 0.45 | 0.8086 |

**The SAS System**

**The GLM Procedure**

| Number of Observations Read | 25 |
|---|---|
| Number of Observations Used | 25 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: percent**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 10 | 442.560000 | 44.256000 | 0.64 | 0.7616 |
| Error | 14 | 974.000000 | 69.571429 | | |
| Corrected Total | 24 | 1416.560000 | | | |

| R-Square | Coeff Var | Root MSE | percent Mean |
|---|---|---|---|
| 0.312419 | 22.39782 | 8.340949 | 37.24000 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| source | 3 | 261.5516667 | 87.1838889 | 1.25 | 0.3282 |
| refinery | 2 | 25.5346713 | 12.7673357 | 0.18 | 0.8343 |
| source*refinery | 5 | 155.4736620 | 31.0947324 | 0.45 | 0.8086 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| source | 3 | 282.6326238 | 94.2108746 | 1.35 | 0.2972 |
| refinery | 2 | 10.7659314 | 5.3829657 | 0.08 | 0.9259 |
| source*refinery | 5 | 155.4736620 | 31.0947324 | 0.45 | 0.8086 |

The code that generated the above output is

```
data tests;
Input refinery source percent;
```

```
cards;
1 1 31
1 1 33
.................
3 4 37
3 4 43
proc glm;
class refinery source;
model percent=refinery source refinery*source;
proc glm;
class source refinery;
model percent=source refinery source*refinery;
run;
```

***(iii) Computational Alternatives.*** Equation (63) for $R(\mu, \alpha, \beta)$ is based upon solving the normal equations for the model (62) by "absorbing" the $\alpha$-equations and solving for $(b-1)$ $\beta$'s. This is the procedure described in detail for the no-interaction model in Section 1d. We pointed out there, without explicit presentation of details, that $R(\mu, \alpha, \beta)$ can also be calculated by solving the normal equations by "absorbing" the $\beta$-equations and solving for $(a-1)$ $\alpha$'s. The calculation of $R(\mu, \alpha, \beta)$ is then as follows. We have that

$$R(\mu, \alpha, \beta) = \sum_{j=1}^{b} n_{.j} \bar{y}_{.j.}^2 + \mathbf{u}' \mathbf{T}^{-1} \mathbf{u} \tag{69}$$

where,

$$\mathbf{T} = \{t_{ii'}\} \quad \text{for} \quad i, i' = 1, 2, \ldots, a-1$$

with

$$t_{ii} = n_{i.} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{.j}},$$

$$t_{ii'} = -\sum_{j=1}^{b} \frac{n_{ij} n_{i'j}}{n_{.j}} \quad \text{for} \quad i \neq i', \tag{70}$$

and

$$\mathbf{u} = \{u_i\} = \left\{ y_{i..} - \sum_{j=1}^{b} n_{ij} \bar{y}_{.j.} \right\} \quad \text{for} \quad i = 1, 2, \ldots, a-1.$$

Table 7.6 involves 3 $\alpha$'s and 4 $\beta$'s. For these data, it is therefore, computationally easier to use (69) instead of (63) for calculating $R(\mu, \alpha, \beta)$ because in (69), $\mathbf{T}$ has order 2 whereas in (63) $\mathbf{C}$ has order 3. The difference in effort here is negligible. However, the choice of procedure might be more crucial if, say, there were many more $\beta$'s than $\alpha$'s or vice versa. (Thus (see Chapter 10), where there may be, say,

**TABLE 7.8 Equivalent Expressions for Sums of Squares in the Analysis of Variance of the Two-Way Classification with Interaction**

| Sum of Squares | d.f.[a] | Method: Absorbing $\alpha$'s (Use When More $\alpha$'s Than $\beta$'s) See (63) for $\mathbf{r'C^{-1}r}$ | Method: Absorbing $\beta$'s (Use When More $\beta$'s Than $\alpha$'s)See (69) for $\mathbf{u'T^{-1}u}$ |
|---|---|---|---|
| ***Fitting* $\alpha$ before $\beta$ (Table 7.7b)** | | | |
| $R(\mu)$ | 1 | $n_{..}\bar{y}_{...}^2$ | $n_{..}\bar{y}_{...}^2$ |
| $R(\alpha\|\mu)$ | $a-1$ | $\sum_i n_{i.}\bar{y}_{i.}^2 - n_{..}\bar{y}_{..}^2$ | $\sum_i n_{i.}\bar{y}_{i.}^2 - n_{..}\bar{y}_{..}^2$ |
| $R(\beta\|\mu,\alpha)$ | $b-1$ | $\mathbf{r'C^{-1}r}$ | $\sum_j n_{.j}\bar{y}_{.j}^2 + \mathbf{u'T^{-1}u} - \sum_i n_{i.}\bar{y}_{i..}^2$ |
| $R(\gamma\|\mu,\alpha,\beta)$ | $s-a-b+1$ | $\sum_i\sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{i..}\bar{y}_{i..}^2 - \mathbf{r'C^{-1}r}$ | $\sum_i\sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{.j}\bar{y}_{.j.}^2 - \mathbf{u'T^{-1}u}$ |
| SSE | $N-s$ | $\sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j n_{ij}\bar{y}_{ij.}^2$ | $\sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j n_{ij}\bar{y}_{ij.}^2$ |
| SST | $N$ | $\sum_i\sum_j\sum_k y_{ijk}^2$ | $\sum_i\sum_j\sum_k y_{ijk}^2$ |
| **Fitting $\beta$ before $\alpha$ (Table 7.7c)** | | | |
| $R(\mu)$ | 1 | $n_{..}\bar{y}_{...}^2$ | $n_{..}\bar{y}_{...}^2$ |
| $R(\beta\|\mu)$ | $b-1$ | $\sum_j n_{.j}\bar{y}_{.j.}^2 - n_{..}\bar{y}_{...}^2$ | $\sum_j n_{.j}\bar{y}_{.j.}^2 - n_{..}\bar{y}_{...}^2$ |
| $R(\alpha\|\mu,\beta)$ | $a-1$ | $\sum_i n_{i.}\bar{y}_{i..}^2 + \mathbf{r'C^{-1}r} - \sum_j n_{.j.}\bar{y}_{.j.}^2$ | $\mathbf{u'T^{-1}u}$ |
| $R(\gamma\|\mu,\alpha,\beta)$ | $s-a-b+1$ | $\sum_i\sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{i..}\bar{y}_{i..}^2 - \mathbf{r'C^{-1}r}$ | $\sum_i\sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{.j}\bar{y}_{.j.}^2 - \mathbf{u'T^{-1}u}$ |
| SSE | $N-s$ | $\sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j n_{ij}\bar{y}_{ij.}^2$ | $\sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j n_{ij}\bar{y}_{ij.}^2$ |
| SST | $N$ | $\sum_i\sum_j\sum_k y_{ijk}^2$ | $\sum_i\sum_j\sum_k y_{ijk}^2$ |

[a] $s$ = number of fitted cells.

2000 $\beta$'s and only 12 $\alpha$'s. Then, (69) requiring inversion of a matrix of order 11, is clearly preferable to (63) which demands inverting a matrix of order 1999!)

The two alternative procedures for calculating $R(\mu, \alpha, \beta)$ provide identical numerical results, but different symbolic expressions for certain of the sums of square in Table 7.7. We show these expressions in Table 7.8 under the headings "Absorbing $\alpha$'s" and "Absorbing $\beta$'s", which describe the method for solving the normal equations implicit in the procedures. For any given set of data, only one of the procedures need to be used. However, the other always provides a check on the arithmetic involved. The choice of which to use depends on whether there are more or fewer $\alpha$'s than $\beta$'s. We can avoid this choice by always denoting the factor with the larger number of effects by $\alpha$. Then, the "Absorbing $\alpha$'s" procedure will be the one to use. Nevertheless, it is of interest to lay out two sets of expressions. We do this in Table 7.8.

(*iv*) ***Interpretation of Results.*** Other than $F(M)$, the $F$-statistics in Tables 7.7b and 7.7c are not significant. We would expect this because $F(R_{\mathrm{m}})$ of Table 7.7a is not significant (see the end of Section 1e). In general, interpretation of the test statistics $F(\alpha|\mu)$, $F(\beta|\mu,\alpha)$, $F(\beta|\mu)$, and $F(\alpha|\mu,\beta)$ in Tables 7.7b and 7.7c is exactly as given in Table 7.4. The possibilities concerning the significance and the non-significance of the $F$'s are the same here and there. Therefore, the interpretation is the same. Furthermore, Tables 7.7b and 7.7 both have the statistic $F(\gamma|\mu,\alpha,\beta)$. This provides a test the effectiveness (in terms of accounting for the variation in $y$) of fitting the model (51) in comparison with fitting the model (1). Since the difference between the two models is the fitting of the interaction effect $\gamma_{ij}$ in (51), we often refer the test as a test for interaction after fitting the main effects. However, like Table 7.2, the interpretation of $F$-statistics can be thought of in two ways:

1. as we have already considered, testing the effectiveness of fitting different models;
2. testing linear hypotheses about elements of the model.

The context of the second interpretation of the $F$-statistics makes us better able to consider the meaning of the tests provided by Table 7.7. First, however, we deal with a limitation of the $R(\ )$ notation and then, in order to discuss tests of hypothesis, consider estimable functions.

(*v*) ***Fitting Main Effects Before Interaction.*** We have defined and freely used notation of the form $R(\mu,\alpha) - R(\mu)$. Formally, it might seem plausible to define

$$R(\beta|\mu,\alpha,\gamma) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\gamma).$$

However, before trying to do this, we must take a careful look at the meaning of the interaction $\gamma$-factor. In doing so, we will find that $R(\beta|\mu,\alpha,\gamma)$, as formally defined by the notation, is identically equal to zero. Evidence of this comes from the models (and corresponding sums of squares) that the notation $R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\gamma)$ implies. For $R(\mu,\alpha,\beta,\gamma)$, the model is (51) and

$$R(\mu,\alpha,\beta,\gamma) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.}^2$$

as in (61). Similarly, in the context of the $\alpha$'s and $\gamma$'s of (51), the implied model for $R(\mu,\alpha,\gamma)$ is $y_{ijk} = \mu + \alpha_i + \gamma_{ij} + e_{ijk}$. However, this is exactly the model for the two-way nested classification discussed in Section 4 of Chapter 6. Hence, the corresponding reduction in the sum of squares is

$$R(\mu,\alpha,\gamma) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.}^2. \tag{71}$$

Consequently,

$$R(\beta|\mu,\alpha,\gamma) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\gamma) \equiv 0.$$

Similarly,

$$R(\mu,\beta,\gamma) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.}^2 = R(\mu,\gamma). \tag{72}$$

Thus, we also have

$$R(\alpha|\mu,\beta,\gamma) \equiv 0 \equiv R(\alpha,\beta|\mu,\gamma).$$

From (61), (71), and (72), we see that the reduction in sum of squares due to fitting any model that contains the interaction $\gamma$-factor is $\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.}^2$. More particularly, in (71) and (72), the reduction of any model which, compared to (51), lacks either $\alpha$, or $\beta$, or both, is equal to $R(\mu,\alpha,\beta,\gamma) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.}^2$. Indeed, as in (72), fitting just ($\mu$ and) the $\gamma$-factor alone leads to the same reduction in the sum of squares. We return to this fact later. Meanwhile, the emphasis here is that in the $R(\ )$ notation, there is no such thing as $R(\beta|\mu,\alpha,\gamma)$ when $\gamma$ is the interaction factor between the $\alpha$- and $\beta$-factors. This is the underlying reason for there being only two subsections of Table 7.7 after 7.7a. There in 7.7b, we have $R(\mu), R(\alpha|\mu), R(\beta|\mu,\alpha),$ and $R(\gamma|\mu,\alpha,\beta)$ based on fitting $\mu,\alpha,\beta,$ and $\gamma$ in that order. Likewise, in Table 7.7c, we have $R(\mu)$,$R(\beta|\mu)$, $R(\alpha|\mu,\beta)$, and $R(\gamma|\mu,\alpha,\beta)$ for fitting $\mu,\beta,\alpha,$ and $\gamma$ in that order. Notationally, one might be tempted from this to consider other sequences such as $\mu,\alpha,\gamma,$ and $\beta$, for example. This would give rise to the notation $R(\mu), R(\alpha|\mu), R(\alpha,\gamma|\mu),$ and $R(\gamma|\mu,\alpha,\beta)$. However, the latter symbol is, as we have seen identically equal to zero, is not a sum of squares. As a result, in the fitting of $\alpha$-, $\beta$-, and $\gamma$-factors with $\gamma$ representing $\alpha$- by $\beta$-interactions, we can fit $\gamma$ only in combination with both $\alpha$ and $\beta$. We cannot fit $\gamma$ unless both $\alpha$ and $\beta$ are in the model. Generally, it is true that in the context of the kind of models being considered here, interaction factors can only be fitted when all their corresponding main effects are in the model too. Moreover, only $R(\ )$ symbols adhering to this policy have meaning.

### e. Estimable Functions

The basic estimable function for the two-way classification model with interaction is

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

We shall frequently refer to this expression. Hence, we give it the symbol $\mu_{ij}$. Thus,

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{73}$$

Its b.l.u.e. is

$$\hat{\mu}_{ij} = \widehat{\mu + \alpha_i + \beta_j + \gamma_{ij}.} = \mu^\circ + \alpha_i^\circ + \beta_j^\circ + \gamma_{ij}^\circ.$$

Since the only non-zero elements in $\mathbf{b}^\circ$ of (55) are $\gamma_{ij}^\circ = \bar{y}_{ij.}$, we have that

$$\hat{\mu}_{ij} = \bar{y}_{ij.}. \tag{74}$$

Moreover,

$$v(\hat{\mu}_{ij}) = \frac{\sigma^2}{n_{ij}}, \tag{75a}$$

and

$$\operatorname{cov}(\hat{\mu}_{ij}, \hat{\mu}_{i'j'}) = 0 \quad \text{unless} \quad i = i' \text{ and } j = j'. \tag{75b}$$

These results are fundamental to the ensuing discussion.

By its definition in (73), $\mu_{ij}$ corresponds to the cell in row $i$ and column $j$ of the grid of rows and columns in which the data may be displayed (e.g., Table 7.6). Therefore, $\mu_{ij}$ is estimable only if the corresponding $(i, j)$ cell contains observations. This also follows from (74) wherein, $\hat{\mu}_{ij}$ the b.l.u.e. of $\mu_{ij}$ exists only for cells that have data, that is, for which there is a $\bar{y}_{ij.}$ value. Therefore, when we say that $\mu_{ij}$ is estimable, we implicitly refer only to those $\mu_{ij}$'s that correspond to cells that have data. The other $\mu_{ij}$'s are not estimable.

Any linear function of the estimable $\mu_{ij}$'s is estimable. However, because of the presence of $\gamma_{ij}$ in $\mu_{ij}$, differences between the rows ($\alpha$'s) or columns ($\beta$'s) are not estimable. For example, in Table 7.6, $\bar{y}_{11.}$ and $\bar{y}_{21.}$ exist. As a result $\mu_{11}$ and $\mu_{21}$ are estimable. Therefore,

$$\mu_{11} - \mu_{21} = \alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21}$$

is estimable. However, $\alpha_1 - \alpha_2$ is not estimable. Similarly, $\alpha_1 - \alpha_3 + \gamma_{13} - \gamma_{33}$ and $\beta_1 - \beta_3 + \gamma_{11} - \gamma_{13}$ are estimable but $\alpha_1 - \alpha_3$ and $\beta_1 - \beta_3$ are not. In general,

$$\alpha_i - \alpha_{i'} + \sum_{j=1}^{b} k_{ij}(\beta_j + \gamma_{ij}) - \sum_{j=1}^{b} k_{i'j}(\beta_j + \gamma_{i'j}) \tag{76}$$

for $i \neq i'$ is estimable provided that

$$\sum_{j=1}^{b} k_{ij} = 1 = \sum_{j=1}^{b} k_{i'j} \tag{77}$$

with $k_{ij} = 0$ when $n_{ij} = 0$ and $k_{i'j} = 0$ when $n_{i'j} = 0$. Then the b.l.u.e. of (76) is

$$\sum_{j=1}^{b} k_{ij}\bar{y}_{ij.} - \sum_{j=1}^{b} k_{i'j},\bar{y}_{i'j}, \tag{78}$$

with variance

$$\sum_{j=1}^{b} \left( \frac{k_{ij}^2}{n_{ij}} + \frac{k_{i'j}^2}{n_{i'j}} \right) \sigma^2.$$

A similar result holds for the $\beta$'s. The parametric function

$$\beta_j - \beta_{j'} + \sum_{i=1}^{a} h_{ij}(\alpha_i + \gamma_{ij}) - \sum_{i=1}^{a} h_{ij'}(\alpha_i + \gamma_{ij'}), \tag{79}$$

is estimable provided that

$$\sum_{i=1}^{a} h_{ij} = 1 = \sum_{i=1}^{a} h_{ij'} \tag{80}$$

where $h_{ij} = 0$ when $n_{ij} = 0$ and $h_{ij'} = 0$ when $n_{ij'} = 0$. The b.l.u.e. of (79) is

$$\sum_{i=1}^{a} h_{ij}\bar{y}_{ij.} - \sum_{i=1}^{a} h_{ij'}\bar{y}_{ij'}. \tag{81}$$

It is not possible to derive an estimable function from the $\mu_{ij}$'s which is solely a function of the $\alpha$'s and $\beta$'s. On the other hand, it is possible to derive an estimable function that is a function of only the $\gamma$'s. Provided that the $(ij), (i'j), (ij')$, and $(i'j')$ cells have data in them, the parametric function

$$\begin{aligned} \theta_{ij,i'j'} &\equiv \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \\ &= \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} \end{aligned} \tag{82}$$

is estimable. Its b.l.u.e. is

$$\hat{\theta}_{ij,i'j'} = \bar{y}_{ij.} - \bar{y}_{i'j.} - \bar{y}_{ij'.} + \bar{y}_{i'j'.}. \tag{83}$$

The variance of (83) is

$$v(\hat{\theta}_{ij,i'j'}) = \left(\frac{1}{n_{ij}} + \frac{1}{n_{i'j}} + \frac{1}{n_{ij'}} + \frac{1}{n_{i'j'}}\right)\sigma^2.$$

Expressions (76), (79), and (82) are the nearest we can come to obtaining estimable functions of intuitively practical value. We cannot estimate differences between row effects devoid of interaction effects. They are not estimable. They can be estimated only in the presence of average column and interaction effects. For example, with $k_{ij} = 1/m_i$, where $m_i$ is the number of filled cells in the $i$th row (i.e., $n_{ij} \neq 0$ for $m_i$ values of $j = 1, 2, \ldots, b$), (77) is satisfied and

$$\alpha_i - \alpha_{i'} + \sum_{\substack{j \text{ for} \\ n_{ij}\neq 0}} \frac{(\beta_j + \gamma_{ij})}{m_i} - \sum_{\substack{j \text{ for} \\ n_{ij'}\neq 0}} \frac{(\beta_j + \gamma_{i'j})}{m_{i'}} \tag{84a}$$

is estimable with b.l.u.e.

$$\sum_{\substack{j \text{ for} \\ n_{ij}\neq 0}} \frac{\bar{y}_{ij.}}{m_i} - \sum_{\substack{j \text{ for} \\ n_{i'j}\neq 0}} \frac{\bar{y}_{i'j.}}{m_{i'}}.$$

Similarly, because $k_{ij} = n_{ij}/n_{i.}$ also satisfies (77), the parametric function

$$\alpha_i - \alpha_{i'} + \sum_{j=1}^{b} \frac{n_{ij}(\beta_j + \gamma_{ij})}{n_{i.}} - \sum_{j=1}^{b} \frac{n_{i'j}(\beta_j + \gamma_{i'j})}{n_{i'.}} \tag{84b}$$

is also estimable. Its b.l.u.e. is

$$\sum_j \frac{n_{ij}\bar{y}_{ij.}}{n_{i.}} - \sum_j \frac{n_{i'j}\bar{y}_{i'j.}}{n_{i'.}} = \bar{y}_{i..} - \bar{y}_{i'..}.$$

**Example 11 Some Estimable Functions and their b.l.u.e.'s** Table 7.6 provides the following examples. First, from (76)–(78)

$$\alpha_1 - \alpha_2 + (\beta_1 + \gamma_{11}) - (\beta_1 + \gamma_{21}) = \alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21} \tag{85a}$$

is estimable with b.l.u.e.

$$\bar{y}_{11.} - \bar{y}_{21.} = 10 - 9 = 1.$$

Similarly,

$$\alpha_1 - \alpha_2 + (\beta_1 + \gamma_{11}) - \tfrac{1}{2}(\beta_1 + \beta_2 + \gamma_{21} + \gamma_{22}) \tag{85b}$$

is estimable with b.l.u.e.

$$\bar{y}_{11.} - \tfrac{1}{2}(\bar{y}_{21.} + \bar{y}_{22.}) = 10 - \tfrac{1}{2}(9 + 13) = -1.$$

As far as $\alpha_1 - \alpha_2$ is concerned, the two estimable functions in (85a) and (85b) are the same. Of course, they involve different functions of the $\beta$'s and the $\gamma$'s. In (85a), there are no $\beta$'s because for both rows (treatments) 1 and 2, there are observations in column (variety) 1. (See Table 7.6). An example of (84b) is that

$$\alpha_1 - \alpha_2 + \frac{[3(\beta_1 + \gamma_{11}) + (\beta_3 + \gamma_{13}) + 2(\beta_4 + \gamma_{14})]}{6} - \frac{[2(\beta_2 + \gamma_{32}) + 2(\beta_3 + \gamma_{33}) + 4(\beta_4 + \gamma_{44})]}{8}$$

is estimable with b.l.u.e.

$$\bar{y}_{1..} - \bar{y}_{3.} = 10 - 11.75 = -1.75. \qquad \square$$

Certain other estimable functions deserve mention because they arise in the discussion of tests of hypotheses corresponding to the $F$-statistics of Table 7.7. The first is

$$\varphi_i = \left(n_{i.} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{.j}}\right)\alpha_i - \sum_{i' \neq i}^{a} \left(\sum_{j=1}^{b} \frac{n_{ij}n_{i'j}}{n_{.j}}\right)\alpha_{i'} + \sum_{j=1}^{b} \left(n_{ij} - \frac{n_{ij}^2}{n_{.j}}\right)\gamma_{ij} - \sum_{i' \neq i}^{a} \left(\sum_{j=1}^{b} \frac{n_{ij}n_{i'j}}{n_{.j}}\right)\gamma_{i'j}. \tag{86}$$

Recall that a linear combination of estimable functions is estimable. The expression in (86) may be rewritten as

$$\varphi_i = \sum_{j=1}^{b} \left[n_{ij}(\mu + \alpha_i + \beta_j + \gamma_{ij}) - \sum_{k=1}^{a} \frac{n_{ij}n_{kj}}{n_{.j}} \left(\mu + \alpha_k + \beta_j + \gamma_{kj}\right)\right].$$

It is an estimable function because it is the linear combination of two basic estimable functions. Another similar expression in terms of $\beta$'s and $\gamma$'s that is also estimable is

$$\psi_j = \left(n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}}\right)\beta_{.j} - \sum_{j' \neq j}^{b} \left(\sum_{i=1}^{a} \frac{n_{ij}n_{ij'}}{n_{i.}}\right)\beta_{j'} + \sum_{i=1}^{a} \left(n_{ij} - \frac{n_{ij}^2}{n_{i.}}\right)\gamma_{ij} - \sum_{j' \neq j}^{b} \left(\sum_{i=1}^{a} \frac{n_{ij}n_{ij'}}{n_{i.}}\gamma_{ij'}\right). \tag{87}$$

Naturally, for $\theta_{ij,i'j'}$ as defined in (82), functions of estimable $\theta$'s are also estimable. However, certain functions of non-estimable $\theta$'s are also estimable. For example, with the data of Table 7.6

$$\theta_{11,22} = \mu_{11} - \mu_{21} - \mu_{12} + \mu_{22} \quad \text{and} \quad \theta_{12,33} = \mu_{12} - \mu_{32} - \mu_{13} + \mu_{33}$$

are not estimable, because $\mu_{12}$ is not. (There are no observations for treatment 1 and variety 2.) However, the sum of these two $\theta$'s does not involve $\mu_{12}$ and is estimable. Thus, we have that

$$\begin{aligned} \delta &= \theta_{11,22} + \theta_{12,33} = \mu_{11} - \mu_{21} + \mu_{22} - \mu_{32} - \mu_{13} + \mu_{33} \\ &= \mu_{11} - \mu_{13} - \mu_{21} + \mu_{22} - \mu_{32} + \mu_{33} \end{aligned} \tag{88}$$

is estimable. For each of the $\mu_{ij}$ in (88), there is at least one observation in treatment $i$ and variety $j$ so that all of the $\mu_{ij}$ are estimable. Hence $\delta$ is estimable. In general, if each of the two non-estimable $\theta$'s involves only a single non-estimable $\mu_{ij}$ which is common to both $\theta$'s, then the sum or difference of those $\theta$'s will not involve that $\mu_{ij}$ and will be estimable. An example of this situation is given by (88).

### f. Tests of Hypotheses

**(*i*) *The General Hypothesis.*** As has already been well-established, the $F$-statistic for testing $H : \mathbf{K}'\mathbf{b} = \mathbf{0}$, is for $\mathbf{K}'$ of full rank $s^*$,

$$F = \frac{Q}{s^*\hat{\sigma}^2} \quad \text{with} \quad Q = (\mathbf{K}'\mathbf{b})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^\circ. \tag{89}$$

Furthermore, hypotheses are testable only when they can be expressed in terms of estimable functions—in this case in terms of the $\mu_{ij}$'s. Thus, any testable hypothesis concerning $\mathbf{K}'\mathbf{b}$ will involve linear functions of the $\mu_{ij}$'s. By the nature of the $\mu_{ij}$, no matter what functions of the $\alpha$'s and $\beta$'s are involved in $\mathbf{K}'\mathbf{b}$, the functions of the $\gamma_{ij}$'s will be the same as those of the $\mu_{ij}$'s. Thus, if

$$\mu = \{\mu_{ij}\} \quad \text{and} \quad \gamma = \{\gamma_{ij}\}, \text{ for } n_{ij} \neq 0, \tag{90}$$

then, when $\mathbf{K}'\mathbf{b} = \mathbf{L}'\mu$, that part of $\mathbf{K}'\mathbf{b}$ that involves $\gamma$ is $\mathbf{L}'\gamma$. In (55), the only non-zero elements of $\mathbf{b}^\circ$ are

$$\gamma^\circ = \{\gamma_{ij}^\circ\} = \overline{\mathbf{y}} = \{\overline{y}_{ij}\} \quad \text{for} \quad n_{ij} \neq 0. \tag{91}$$

Similarly, in (57), the only non-null sub-matrix in $\mathbf{G}$ is the diagonal matrix $\mathbf{D}\{\frac{1}{n_{ij}}\}$ corresponding to $\gamma^\circ$. Therefore, to test the hypothesis

$$H : \mathbf{K}'\mathbf{b} = \mathbf{0} \text{ equivalent to } \mathbf{L}'\mu = \mathbf{0}, \tag{92}$$

$Q$ of (89) becomes

$$Q = \bar{\mathbf{y}}'\mathbf{L}\left[\mathbf{L}'\mathbf{D}\left\{\frac{1}{n_{ij}}\right\}\mathbf{L}\right]^{-1}\mathbf{L}\bar{\mathbf{y}}. \tag{93}$$

**Example 12 Testing the Hypothesis $\mathbf{L}'\boldsymbol{\mu} = \mathbf{0}$ for Data of Table 7.6** For Table 7.6,

$$\gamma' = \begin{bmatrix}\gamma_{11} & \gamma_{13} & \gamma_{14} & \gamma_{21} & \gamma_{22} & \gamma_{32} & \gamma_{33} & \gamma_{34}\end{bmatrix} \tag{94}$$

and

$$\begin{aligned}\bar{\mathbf{y}}' &= \begin{bmatrix}\bar{y}_{11.} & \bar{y}_{13.} & \bar{y}_{14.} & \bar{y}_{21.} & \bar{y}_{22.} & \bar{y}_{32.} & \bar{y}_{33.} & \bar{y}_{34.}\end{bmatrix} \\ &= \begin{bmatrix}10 & 12 & 9 & 9 & 13 & 8 & 15 & 12\end{bmatrix}.\end{aligned} \tag{95}$$

In (85), $\alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21}$ is estimable, for which $\mathbf{L}'\gamma = \gamma_{11} - \gamma_{12}$ has

$$\mathbf{L}' = \begin{bmatrix}1 & 0 & 0 & -1 & 0 & 0 & 0 & 0\end{bmatrix}.$$

In addition, from (53),

$$\mathbf{D} = \mathbf{D}\left\{\frac{1}{n_{ij}}\right\} = \operatorname{diag}\left[\frac{1}{3} \quad 1 \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{4}\right], \tag{96}$$

so that

$$\mathbf{L}'\mathbf{D} = \begin{bmatrix}\frac{1}{3} & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0\end{bmatrix}$$

and

$$\mathbf{L}'\mathbf{D}\mathbf{L} = \left(\frac{1}{3} + \frac{1}{2}\right) = \frac{5}{6}.$$

Therefore, for testing the hypothesis $\alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21} = 0$ we have

$$Q = (10 - 9)\left(\frac{5}{6}\right)^{-1}(10 - 9) = 1.2.$$

In this way, we need only look at the $\gamma_{ij}$ elements of a hypothesis in order to derive $\mathbf{L}'$ and so calculate $Q$ of (89). □

**(*ii*) *The Hypothesis for F(M).*** In earlier discussing Table 7.7, we interpreted the sum of squares therein as reductions in the sum of squares due to fitting different models. We now consider their meaning in terms of testing hypotheses. In this context, we only deal with hypotheses about the two-way classification interaction model, (51). In particular, we establish the linear hypotheses corresponding to each

of the six different $F$-statistics in Tables 7.7b and 7.7c. The first of these $F$-statistics is $F(M)$.

Results for the general case (e.g., equation (15) of Chapter 6) indicate that $F(M)$ can be used to test the hypothesis $H$: $E(\overline{y}) = 0$. In the present case, this is equivalent to

$$H\colon \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\mu_{ij} = 0 \quad \text{for} \quad n_{ij} \neq 0. \tag{97}$$

In terms of (92), the hypothesis in (97) can be expressed as $\mathbf{L}'\boldsymbol{\mu} = \mathbf{0}$ for $\mathbf{L}'$ being the vector

$$\mathbf{L}'\mathbf{L}' = \begin{bmatrix} n_{11} & \cdots & n_{ab} \end{bmatrix} \quad \text{for those} \quad n_{ij} \neq 0. \tag{98}$$

Hence,

$$\mathbf{L}'\mathbf{D} = \mathbf{1}' \quad \text{with} \quad \mathbf{L}'\mathbf{D}\mathbf{L} = N. \tag{99}$$

For (93), we can show that $\mathbf{L}'\overline{\mathbf{y}} = y_{\cdots}$. As a result, (93) becomes $Q = R(\boldsymbol{\mu})$. This confirms the numerator of $F(M)$.

***(iii) Hypotheses for $F(\boldsymbol{\alpha}|\boldsymbol{\mu})$ and $F(\boldsymbol{\beta}|\boldsymbol{\mu})$.*** We will show that $R(\boldsymbol{\alpha}|\boldsymbol{\mu})$ is the numerator sum of squares for testing

$$H\colon \frac{1}{n_{i\cdot}}\sum_{i} n_{ij}\boldsymbol{\mu}_{ij} \text{ equal for all } i. \tag{100a}$$

This hypothesis may also be stated as

$$H\colon \boldsymbol{\alpha}_i + \frac{1}{n_{i\cdot}}\sum_{j} n_{ij}(\boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}) \text{ equal for all } i. \tag{100b}$$

The hypothesis in (100a) and (100b) can be expressed as $a - 1$ independent differences

$$H\colon \frac{1}{n_{1\cdot}}\sum_{j} n_{1j}\mu_{1j} - \frac{1}{n_{i\cdot}}\sum_{j} n_{ij}\mu_{ij} = 0 \quad \text{for} \quad i = 2, 3, \ldots, a. \tag{100c}$$

We can then see that for (93) that the $(i-1)$th row of $\mathbf{L}'$ is for $i = 2, 3, \ldots, a$,

$$\boldsymbol{\ell}'_{i-1} = \Bigg[ \underbrace{\frac{n_{11}}{n_{1\cdot}} \quad \cdots \quad \frac{n_{1b}}{n_{1\cdot}}}_{\substack{\text{corresponding} \\ \text{to } n_{1j} \neq 0}} \quad \mathbf{0}' \quad \underbrace{-\frac{n_{i1}}{n_{i\cdot}} \quad \cdots \quad -\frac{n_{ib}}{n_{i\cdot}}}_{\substack{\text{corresponding} \\ \text{to } n_{ij} \neq 0}} \quad \mathbf{0}' \Bigg]. \tag{101}$$

We can show from this that the $(i-1)$th element of $\mathbf{L}'\bar{\mathbf{y}}$ is $\bar{y}_{1..} - \bar{y}_{i..}$ and that $\mathbf{L}'\mathbf{D}\mathbf{L} = (1/n_{1.})\mathbf{J} + \mathbf{D}\{n_{i.}\}$ for $i = 2, 3, \ldots, a$. Algebraic simplification based on results in Exercise 20 leads to $Q$ of (93) reducing to $R(\boldsymbol{\alpha}|\boldsymbol{\mu})$. Hence (100b) is the hypothesis tested by $F(\boldsymbol{\alpha}|\boldsymbol{\mu})$.

**Example 13 The Numerator of the Test Statistic for $R(\boldsymbol{\alpha}|\boldsymbol{\mu})$ in Table 7.6** For the data of Table 7.6, consider

$$
\begin{aligned}
H\colon\ & \alpha_1 + \tfrac{1}{6}[3(\beta_1 + \gamma_{11}) + (\beta_3 + \gamma_{13}) + 2(\beta_4 + \gamma_{14})] \\
& -\{\alpha_2 + \tfrac{1}{4}[2(\beta_1 + \gamma_{21}) + 2(\beta_2 + \gamma_{22})]\} = 0 \\
& \alpha_1 + \tfrac{1}{6}[3(\beta_1 + \gamma_{11}) + (\beta_3 + \gamma_{13}) + 2(\beta_4 + \gamma_{14})] \\
& -\{\alpha_3 + \tfrac{1}{8}[2(\beta_2 + \gamma_{32}) + 2(\beta_3 + \gamma_{33}) + 4(\beta_4 + \gamma_{34})]\} = 0.
\end{aligned}
$$

We then have,

$$
\mathbf{L}' = \begin{bmatrix} \frac{3}{6} & \frac{1}{6} & \frac{2}{6} & -\frac{2}{4} & -\frac{2}{4} & 0 & 0 & 0 \\ \frac{3}{6} & \frac{1}{6} & \frac{2}{6} & 0 & 0 & -\frac{2}{8} & -\frac{2}{8} & -\frac{4}{8} \end{bmatrix}, \tag{102}
$$

$$
\mathbf{L}'\bar{\mathbf{y}} = \begin{bmatrix} \bar{y}_{1..} - \bar{y}_{2..} \\ \bar{y}_{1..} - \bar{y}_{3..} \end{bmatrix} = \begin{bmatrix} 10 - 11 \\ 10 - 11.75 \end{bmatrix} = \begin{bmatrix} -1 \\ -1.75 \end{bmatrix}
$$

and

$$
\mathbf{L}'\mathbf{D} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \end{bmatrix} \tag{103}
$$

where $\mathbf{D}$ is given by (96). Furthermore,

$$
\mathbf{L}'\mathbf{D}\mathbf{L} = \begin{bmatrix} \frac{5}{12} & \frac{1}{6} \\ \frac{1}{6} & \frac{7}{24} \end{bmatrix} \quad \text{with } (\mathbf{L}'\mathbf{D}\mathbf{L})^{-1} = \frac{4}{9}\begin{bmatrix} 7 & -4 \\ -4 & 10 \end{bmatrix}.
$$

Hence,

$$
Q = \begin{bmatrix} -1 & -1.75 \end{bmatrix} \frac{4}{9} \begin{bmatrix} 7 & -4 \\ -4 & 10 \end{bmatrix} \begin{bmatrix} -1 \\ -1.75 \end{bmatrix} = 10.5 = R(\alpha|\mu)
$$

of Table 7.7b. □

Analogous to the above, $R(\beta|\mu)$ is the numerator sum of squares for testing

$$H\colon \beta_j + \frac{1}{n_{\cdot j}} \sum_i n_{ij}(\alpha_i + \gamma_{ij}) \text{ equal for all } j.$$

Exercise 11 provides an example.

**(*iv*) *Hypotheses for $F(\alpha|\mu, \beta)$ and $F(\beta|\mu, \alpha)$.*** These $F$-statistics test the hypotheses

$$H\colon \varphi_i = 0 \text{ for all } i \quad \text{and} \quad H\colon \psi_j = 0 \text{ for all } j,$$

respectively, where $\varphi_i$ and $\psi_j$ are given by (86) and (87). First, observe that $\sum_{i=1}^{a} \varphi_i = 0$. To see this, notice that

$$\begin{aligned}
\sum_{i=1}^{a} \varphi_i &= \sum_{i=1}^{a} \alpha_i \left[ n_{i\cdot} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{\cdot j}} - \sum_{i' \neq i}^{a} \sum_{j=1}^{b} \frac{n_{ij} n_{i'j}}{n_{\cdot j}} \right] + \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} \left[ n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}} - \sum_{i' \neq i} \frac{n_{ij} n_{i'j}}{n_{\cdot j}} \right] \\
&= \sum_{i=1}^{a} \alpha_i \left[ n_{i\cdot} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{\cdot j}} - \sum_{j=1}^{b} \frac{n_{ij}(n_{\cdot j} - n_{ij})}{n_{\cdot j}} \right] + \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} \left[ n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}} - \frac{n_{ij}(n_{\cdot j} - n_{i'j})}{n_{\cdot j}} \right] \\
&\equiv 0.
\end{aligned}$$

Therefore, the hypotheses in $H\colon \varphi_i = 0$ for all $i$ are not independent. Restating them as a set of independent hypotheses, we have

$$H\colon \varphi_i = 0 \quad \text{for} \quad i = 1, 2, \ldots, a-1. \tag{104}$$

Writing these hypotheses in the form $\mathbf{L}'\mu = 0$ the $i$th row of $\mathbf{L}'$ is, for $i = 1, \ldots, a-1$, given by

$$\ell_i' = \left[ \begin{array}{l}
\left\{ -\dfrac{n_{ij} n_{kj}}{n_{\cdot j}} \right\} \text{ for } j = 1, \ldots, b \text{ and } k = 1, \ldots, i-1, \text{ and } n_{kj} \neq 0 \\
\left\{ n_{ij} - \dfrac{n_{ij}^2}{n_{\cdot j}} \right\} \text{ for } j = 1, \ldots, b \text{ and } n_{ij} \neq 0 \\
\left\{ -\dfrac{n_{ij} n_{kj}}{n_{\cdot j}} \right\} \text{ for } j = 1, \ldots, b \text{ and } k = i+1, \ldots, a, \text{ and } n_{kj} \neq 0
\end{array} \right]'. \tag{105}$$

We may then show that, for (93), the $i$th element of $\mathbf{L}'\overline{\mathbf{y}}$ is

$$\sum_{k\neq j}\sum_{j}\left(-\frac{n_{ij}n_{kj}}{n_{.j}}\right)\overline{y}_{kj}+\sum_{j}\left(n_{ij}-\frac{n_{ij}^2}{n_{.j}}\right)\overline{y}_{ij.}=y_{i..}-\sum_{j}n_{ij}\overline{y}_{.j.}.$$

Thus,

$$\mathbf{L}'\overline{\mathbf{y}}=\left\{y_{i..}-\sum_{j=1}^{b}n_{ij}\overline{y}_{.j.}\right\}\quad\text{for}\quad i=1,\ldots,a-1.$$

In a like manner, we can show that the diagonal elements of $\mathbf{L}'\mathbf{D}\mathbf{L}$ are

$$n_{i.}-\sum_{j=1}^{b}\frac{n_{ij}^2}{n_{.j}}\quad\text{for}\quad i=1,2,\ldots,a-1$$

and that the off-diagonal elements of $\mathbf{L}'\mathbf{D}\mathbf{L}$ are

$$-\sum_{j=1}^{b}\frac{n_{ij}n_{i'j}}{n_{.j}}\quad\text{for}\quad i\neq i'=1,2,\ldots,a-1.$$

By analogy for testing

$$H\colon \psi_j=0\quad\text{for}\quad j=1,\ldots,b-1$$

we have,

$$\mathbf{L}'\overline{\mathbf{y}}=\left\{y_{.j.}-\sum_{i=1}^{a}n_{ij}\overline{y}_{i..}\right\}\quad\text{for}\quad j=1,2,\ldots,b-1.$$

The matrix $\mathbf{L}'\mathbf{D}\mathbf{L}$ has diagonal elements

$$n_{.j}-\sum_{i=1}^{a}\frac{n_{ij}^2}{n_{i.}}\quad\text{for}\quad j=1,2,\ldots,b-1$$

and off-diagonal elements

$$-\sum_{i=1}^{a}\frac{n_{ij}n_{ij'}}{n_{i.}}\quad\text{for}\quad j\neq j'=1,2,\ldots,b-1.$$

However, in this case, we see from (64) and (65) that

$$\mathbf{L}'\overline{\mathbf{y}} = \mathbf{r} \quad \text{and} \quad \mathbf{L}'\mathbf{D}\mathbf{L} = \mathbf{C}.$$

Therefore, in (93), from (63)

$$\mathbf{Q} = \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} = R(\beta|\mu, \alpha).$$

Hence, $F(\beta|\mu, \alpha)$ of Table 7.7b tests $H : \psi' = 0$, that is,

$$H : \left(n_{\cdot j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i\cdot}}\right)\beta_j - \sum_{j\neq j'}^{b}\left(\sum_{i=1}^{a} \frac{n_{ij}n_{ij'}}{n_{i\cdot}}\right)\beta_{j'} + \sum_{i=1}^{a}\left(n_{ij} - \frac{n_{ij}^2}{n_{i\cdot}}\right)\gamma_{ij} - \sum_{j\neq j'}^{b}\sum_{i=1}^{a}\left(\frac{n_{ij}n_{ij'}}{n_{i\cdot}}\right)\gamma_{ij'} = 0$$
$$\text{for } j = 1, 2, \ldots, b-1, \tag{106}$$

equivalent to the same hypothesis for $i = 1, 2, \ldots, a$. Correspondingly, $F(\alpha|\mu, \beta)$ of Table 7.7c tests $H$: $\varphi_i = 0$, that is,

$$H\text{: } \left(n_{i\cdot} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{\cdot j}}\right)\alpha_i - \sum_{i'\neq i}^{a}\left(\sum_{j=1}^{b} \frac{n_{ij}n_{i'j}}{n_{\cdot j}}\right)\alpha_{i'} + \sum_{j=1}^{b}\left(n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}}\right)\gamma_{ij} - \sum_{i'\neq i}^{a}\sum_{j=1}^{b}\left(\frac{n_{ij}n_{i'j}}{n_{\cdot j}}\right)\gamma_{i'j} = 0$$
$$\text{for } i = 1, 2, \ldots, a-1. \tag{107}$$

The hypothesis in (107) is equivalent to the same hypothesis for $i = 1, 2, \ldots, a$. Observe that in (106), the coefficients of the $\beta$'s are the elements $c_{jj'}$ of $\mathbf{C}$ in (64) and the coefficients of the $\gamma$'s, if summed over $i$, are also the $c_{jj'}$'s. Analogous properties hold for the coefficients of the $\alpha$'s and the $\gamma$'s in (107).

**Example 14 Calculating the Test Statistics for Testing Hypothesis for $F(\alpha|\mu, \beta)$** According to (92), the $\mathbf{L}'$ matrix for the hypothesis in (107) is obtained for the coefficients of the $\gamma$'s whose terms are

$$\sum_{j=1}^{b}\left(n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}}\right)\gamma_{ij} - \sum_{i'\neq i}^{a}\sum_{j=1}^{b}\left(\frac{n_{ij}n_{i'j}}{n_{\cdot j}}\right)\gamma_{i'j} \quad \text{for} \quad i = 1, 2, \ldots, a-1.$$

For the data of Tables 7.6 and 7.6a, the value of $\mathbf{L}'$ for the hypothesis (107) is

$$\mathbf{L}' = \begin{bmatrix} 3 - \frac{3^2}{5} & 1 - \frac{1^2}{3} & 2 - \frac{2^2}{6} & -\frac{3(2)}{5} & \frac{0(2)}{4} & \frac{0(2)}{4} & -\frac{1(2)}{3} & -\frac{2(4)}{6} \\ -\frac{2(3)}{5} & 0 & 0 & 2 - \frac{2^2}{5} & 2 - \frac{2^2}{4} & -\frac{2^2}{4} & 0 & 0 \end{bmatrix} \tag{108}$$
$$= \frac{1}{15}\begin{bmatrix} 18 & 10 & 20 & -18 & 0 & 0 & -10 & -20 \\ -18 & 0 & 0 & 18 & 15 & -15 & 0 & 0 \end{bmatrix}.$$

Thus,

$$\mathbf{L}'\mathbf{D} = \frac{1}{15}\begin{bmatrix} 6 & 10 & 10 & -9 & 0 & 0 & -5 & -5 \\ -6 & 0 & 0 & 9 & 7.5 & -7.5 & 0 & 0 \end{bmatrix} \tag{109}$$

As a result,

$$\mathbf{L}'\mathbf{D}\mathbf{L} = \frac{1}{5}\begin{bmatrix} 16 & -6 \\ -6 & 11 \end{bmatrix} \quad \text{with} \quad (\mathbf{L}'\mathbf{D}\mathbf{L})^{-1} = \frac{1}{28}\begin{bmatrix} 11 & 6 \\ 6 & 16 \end{bmatrix}.$$

Furthermore,

$$\mathbf{L}'\bar{\mathbf{y}} = \frac{1}{5}\begin{bmatrix} -24 \\ 19 \end{bmatrix}.$$

Therefore, in (93),

$$Q = \frac{1}{5}\begin{bmatrix} -24 & 19 \end{bmatrix}\frac{1}{28}\begin{bmatrix} 11 & 6 \\ 6 & 16 \end{bmatrix}\frac{1}{5}\begin{bmatrix} -24 \\ 9 \end{bmatrix} = 9.49 = R(\alpha|\mu,\beta) \text{ of Table 7.7c.}$$

□

**(*v*) *Hypotheses for $F(\gamma|\mu,\alpha,\beta)$.*** The hypothesis tested by $F(\gamma|\mu,\alpha,\beta)$ takes the following form, where $s-a-b+1$ is the degrees of freedom of the numerator of the $F$-statistic, $R(\gamma|\mu,\alpha,\beta)$.

$$H: \begin{pmatrix} \text{Any column vector consisting of } s-a-b+1 \\ \text{linearly independent function of the} \\ \theta_{ij,i'j'} = \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} \text{ where such} \\ \text{functions are either estimable } \theta'\text{s or estimable} \\ \text{sums or differences of } \theta'\text{s.} \end{pmatrix} = \mathbf{0}. \tag{110}$$

As used here $\theta_{ij,i'j'}$ is as defined in (82). The estimable sums and differences of $\theta'$s are those defined in (88). Writing the hypotheses as

$$\mathbf{L}'\gamma = \mathbf{0},$$

where $\mathbf{L}'$ has order $s-a-b+1$ by $s$ and rank $s-a-b+1$, it follows from the nature of the $\theta'$s that $\mathbf{L}'\mathbf{1} = \mathbf{0}$. Furthermore, the equations $\mathbf{L}'\gamma = \mathbf{0}$ have a solution $\gamma_{ij} = \gamma$ for all $i$ and $j$ for which $n_{ij} \neq 0$. Therefore, the reduced model corresponding to the hypothesis is $E(y_{ijk}) = (\mu+\gamma) + \alpha_i + \beta_j$ for which the reduction in the sum of squares is $R(\mu,\alpha,\beta)$. Therefore, in accord with Section 6d(ii) of Chapter 3,

$$Q = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta) = R(\gamma|\mu,\alpha,\beta).$$

As a result, $F(\gamma|\mu,\alpha,\beta)$ tests the hypothesis in (110).

**Example 15 Test of a Specific Hypothesis About the $\gamma$'s** For the data of Table 7.6, we can test the hypothesis

$$H: \begin{cases} \mu_{13} - \mu_{33} - \mu_{14} + \mu_{34} = 0 \\ \mu_{11} - \mu_{21} - \mu_{12} + \mu_{22} + (\mu_{12} - \mu_{32} - \mu_{13} + \mu_{33}) = 0. \end{cases} \tag{111}$$

In keeping with (82), the first relationship in (111) is $\theta_{13,34} = 0$. As in (88), the second relationship in (111) is $\theta_{12,34} = 0$. Rewriting (111) in terms of the elements of the model, this hypothesis is

$$H: \begin{cases} \gamma_{13} - \gamma_{33} - \gamma_{14} + \gamma_{34} = 0 \\ \gamma_{11} - \gamma_{21} + \gamma_{22} - \gamma_{13} - \gamma_{32} + \gamma_{33} = 0. \end{cases} \tag{112}$$

The second function of the $\gamma$'s in (112) is (88). Writing the hypothesis in (112) as $\mathbf{L}'\gamma = \mathbf{0}$ gives

$$\mathbf{L}' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & -1 & 1 & -1 & 1 & 0 \end{bmatrix}. \tag{113}$$

Then,

$$\mathbf{L}'\overline{\mathbf{y}} = \begin{bmatrix} 0 \\ 9 \end{bmatrix}$$

and

$$\mathbf{L}'\mathbf{D} = \begin{bmatrix} 0 & 1 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & -1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}. \tag{114}$$

As a result,

$$\mathbf{L}'\mathbf{D}\mathbf{L} = \frac{1}{12}\begin{bmatrix} 27 & -18 \\ -18 & 40 \end{bmatrix} \quad \text{and} \quad (\mathbf{L}'\mathbf{D}\mathbf{L})^{-1} = \frac{1}{63}\begin{bmatrix} 40 & 18 \\ 18 & 27 \end{bmatrix}.$$

Hence, in (93),

$$Q = \begin{bmatrix} 0 & 9 \end{bmatrix}\frac{1}{63}\begin{bmatrix} 40 & 18 \\ 18 & 27 \end{bmatrix}\begin{bmatrix} 0 \\ 9 \end{bmatrix} = 34.71 = R(\gamma|\mu, \alpha, \beta)$$

of Tables 7.7b and 7.7c. Hence (111) is the hypothesis tested by $F(\gamma|\mu, \alpha, \beta)$. □

Note that hypotheses of this nature involve not only functions of the form $\theta_{ij,i'j'} = \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'}$, as is the first in (111). They also involve sums and differences of such functions, as is the second in (111). As has already been

explained in the description of $\delta$ in (88), we choose these sums and differences to eliminate a $\mu_{ij}$ that is not estimable. Thus, the second function in (111) is not only

$$\gamma_{11} - \gamma_{21} + \gamma_{22} - \gamma_{13} - \gamma_{32} + \gamma_{33} = \mu_{11} - \mu_{21} - \mu_{12} + \mu_{22} + (\mu_{12} - \mu_{32} - \mu_{13} + \mu_{33}). \tag{115}$$

It is also,

$$\gamma_{11} - \gamma_{21} + \gamma_{22} - \gamma_{13} - \gamma_{32} + \gamma_{33} = \mu_{22} - \mu_{32} - \mu_{23} + \mu_{33} - (\mu_{21} - \mu_{11} - \mu_{23} + \mu_{13}). \tag{116}$$

Equation (116) eliminates the non-estimable $\mu_{23}$. The exact form of these functions corresponding to any particular $R(\gamma|\mu, \alpha, \beta)$ also depends entirely on the available data. The pattern of non-empty cells is the determining factor in establishing which functions of the $\gamma$'s make up the hypotheses tested by $F(\gamma|\mu, \alpha, \beta)$ and which do not. For example, for Table 7.6, one function that is not estimable is

$$\mu_{11} - \mu_{31} - \mu_{13} + \mu_{33} - (\mu_{21} - \mu_{31} - \mu_{24} + \mu_{34}) = \mu_{11} - \mu_{13} + \mu_{33} - \mu_{21} + \mu_{24} - \mu_{34}.$$

This function is not estimable because although it eliminates the non-estimable $\mu_{31}$, it still retains the non-estimable $\mu_{24}$.

**(*vi*) *Reduction to the No-Interaction Model.*** We would anticipate that the hypotheses tested by $F(\alpha|\mu)$ and $F(\beta|\mu, \alpha)$ in the interaction model reduce to those tested by the same statistics in the no-interaction model. We establish this result here.

In the interaction model, $F(\alpha|\mu)$ tests the hypothesis (100b):

$$H\colon \boldsymbol{\alpha}_i + \frac{1}{n_{i.}} \sum_j n_{ij}(\boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}) \text{ equal for all } i.$$

Putting all $\gamma_{ij} = 0$ converts the interaction model into the no-interaction model and transforms the above hypothesis into

$$H\colon \boldsymbol{\alpha}_i + \frac{1}{n_{i.}} \sum_j n_{ij}\boldsymbol{\beta}_j \text{ equal for all } i.$$

This is identical to that tested by $F(\alpha|\mu)$ in the no-interaction model as discussed after Example 6. Similarly, in the interaction model, the hypothesis tested by $F(\beta|\mu, \alpha)$ is that given in (106). Putting all $\gamma_{ij} = 0$ in (106) reduces the hypothesis to

$$H\colon \left(n_{.j} - \sum_{i=1}^{a} \frac{n_{ij}^2}{n_{i.}}\right)\beta_j - \sum_{j \neq j'}^{b} \left(\sum_{i=1}^{a} \frac{n_{ij}n_{ij'}}{n_{i.}}\right)\beta_{j'} = 0 \quad \text{for all } j. \qquad \square$$

This represents $b-1$ linearly dependent equations in $b$ parameters $\beta_1, \beta_2, \ldots, \beta_b$. They only hold when all of the $\beta_j$'s are equal. In this situation, the hypotheses in (106) reduces to $H$: equality of all $\beta_j$'s. This is the hypothesis tested by $F(\beta|\mu, \alpha)$ in the no-interaction model as indicated in Section 1g.

**(*vii*) *Independence Properties.*** As indicated in Section 5g of Chapter 5, the sums of squares for testing hypotheses $\mathbf{k'}_i\mathbf{b} = 0$ and $\mathbf{k'}_j\mathbf{b} = 0$ are, on the basis of underlying normality assumptions, independent if $\mathbf{k'}_i\mathbf{G}\mathbf{k'}_j = 0$. This property can be used to show that the sums of squares in Tables 7.7a, b, and c are independent. To see this, consider $\ell_i'\mathbf{D}\ell_j^*$ where $\ell_i'$ is a row of $\mathbf{L'}$ for one sum of squares and $\ell_j^{*'}$ is a row of $\mathbf{L'}$ for some other sum of squares in the same section of Table 7.7. For example, from (99), $\ell_i'\mathbf{D}$ of $R(\mu)$ is $1'$ and from (102) an $\ell_j^{*'}$ of $R(\alpha|\mu)$ is

$$\ell_j^{*'} = \left[\begin{array}{cccccccc} \frac{3}{6} & \frac{1}{6} & \frac{2}{6} & -\frac{2}{4} & -\frac{2}{4} & 0 & 0 & 0 \end{array}\right].$$

Hence, $\ell_i'\mathbf{D}\ell_j^* = \mathbf{1'}\ell_j^* = 0$.

We will find that the same result is true for the other row of $\mathbf{L'}$ in (102). We thus conclude that $R(\mu)$ and $R(\alpha|\mu)$ are independently distributed. In this way, the independence of the $R(\ )$'s is readily established for Tables 7.7a, b, and c. Expressions for $\mathbf{L'D}$ are given in equations (99), (103), (109), and (114), and for $\mathbf{L'}$ in (98), (102), (108), and (113).

## g. Models that Include Restrictions

Since, as in (76),

$$\alpha_i - \alpha_{i'} + \sum_{j=1}^{b} k_{ij}(\beta_j + \gamma_{ij}) - \sum_{j=1}^{b} k_{i'j}(\beta_j + \gamma_{i'j})$$

is estimable, for the $k$'s satisfying (77), it follows that if the model includes restrictions

$$\sum_{j=1}^{k} k_{ij}(\beta_j + \gamma_{ij}) = 0 \quad \text{for all } i, \quad \text{for} \quad n_{ij} \neq 0, \tag{117}$$

it follows that $\alpha_i - \alpha_{i'}$ is estimable. A particular case of this might be when $k_{ij} = n_{ij}/n_{i.}$, as in (84). In this case, (117) becomes

$$\sum_{j=1}^{k} n_{ij}(\beta_j + \gamma_{ij}) = 0 \quad \text{for all } i, \quad \text{for} \quad n_{ij} \neq 0. \tag{118}$$

Then the corresponding b.l.u.e. of $\alpha_i - \alpha_{i'}$ is then $\bar{y}_{i..} - \bar{y}_{i'..}$. However, there seems to be little merit in having either (117) or (118) as part of a model because both of them are data-dependent. The same thing applies to restrictions that reduce the $F$-statistics of Table 7.7 to hypotheses that have meaningful interpretation, for example, a hypothesis of equality of the $\alpha$'s. As inherent parts of a model, these restrictions suffer from the same deficiencies, as do all such restrictions, as discussed in Sections 1h and 2h of Chapter 6.

### h. All Cells Filled

For data having empty cells such as those of Table 7.7, the nature of which functions are estimable depends entirely on which $n_{ij}$'s are not zero. For example, with the Table 7.6 data, $\alpha_2 + \gamma_{22} - (\alpha_3 + \gamma_{32})$ is estimable but $\alpha_1 + \gamma_{12} - (\alpha_3 + \gamma_{32})$ is not. In contrast, when all cells contain at least one observation,

$$\eta_{ii'} = \alpha_i - \alpha_{i'} + \frac{\left(\sum_{j=1}^{b} \gamma_{ij} - \sum_{j=1}^{b} \gamma_{i'j}\right)}{b} \tag{119}$$

is estimable for all $i \neq i'$. The function in (119) is a special case of that in (76), where $k_{ij} = k_{i'j} = 1/b$. Its b.l.u.e. is

$$\hat{\eta}_{ii'} = \frac{\left(\sum_{j=1}^{b} \bar{y}_{ij.} - \sum_{j=1}^{b} \bar{y}_{i'j.}\right)}{b}. \tag{120}$$

We can test hypotheses about (119). The statistic that tests

$$H\colon \alpha_i - \alpha_{i'} + \frac{\left(\sum_{j=1}^{b} \gamma_{ij} - \sum_{j=1}^{b} \gamma_{i'j}\right)}{b} = 0$$

is

$$F = \frac{\left[\sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i'j.})\right]^2}{\sum_{j=1}^{b} (1/n_{ij} + 1/n_{i'j})\hat{\sigma}^2}$$

with 1 and $s - a - b + 1$ degrees of freedom. Furthermore, we can also test the joint hypothesis

$$H\colon \alpha_i + \sum_{j=1}^{b} \frac{\gamma_{ij}}{b} \text{ all equal, for } i = 1, \ldots, a. \tag{121}$$

The $F$-statistic for doing so is

$$F = \frac{\displaystyle\sum_{i=1}^{a} \frac{\left(\displaystyle\sum_{j=1}^{b} y_{ij.}\right)^2}{\displaystyle\sum_{j=1}^{b} 1/n_{ij}} - \left(\sum_{i=1}^{a} \frac{\displaystyle\sum_{j=1}^{b} y_{ij.}}{\displaystyle\sum_{j=1}^{b} 1/n_{ij}}\right) \Big/ \sum_{i=1}^{a} \frac{1}{\displaystyle\sum_{j=1}^{b} 1/n_{ij}}}{(a-1)\hat{\sigma}^2}. \tag{122}$$

If the model includes the restrictions $\sum_{j=1}^{b} y_{ij.} = 0$ for all $i = 1, 2, \ldots, a$, then (119) reduces to $\alpha_i - \alpha_{i'}$. It is estimable with b.l.u.e. given by (120). The hypothesis (121) becomes $H$: equality of all $\alpha_i$'s. We then test this hypothesis using (122). We can obtain results paralleling (119) through (122) for $\beta$'s in a similar fashion.

### i. Balanced Data

There is great simplification of the preceding results when $n_{ij} = n$ for all $j$, *just as in the no-interaction case and I*. The calculations become those of the familiar two-factor analysis with replication (e.g., see p. 110 of Scheffe (1959) and pp. 255–256 of Gruber (2014)). As before, the solutions to the normal equations are $\gamma_{ij}^{\circ} = \bar{y}_{ij.}$. These are the only non-zero elements of $\mathbf{b}^{\circ}$. If the model includes restrictions $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} \gamma_{ij} = 0$ for all $j$ and $\sum_{j=1}^{b} \gamma_{ij} = 0$ for all $i$, then both $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$ are estimable. Their respective b.l.u.e.'s are

$$\widehat{\alpha_i - \alpha_{i'}} = \bar{y}_{i..} - \bar{y}_{i'..} \quad \text{and} \quad \widehat{\beta_j - \beta_{j'}} = \bar{y}_{.j.} - \bar{y}_{.j'.}.$$

Their respective variances are $2\sigma^2/bn$ and $2\sigma^2/an$. The analysis of variance tables of Table 7.7 and 7.8 also simplify, just as did Tables 7.2 and 7.3 in the no-interaction case. Thus, $R(\alpha|\mu)$ and $R(\alpha|\mu, \beta)$. become identical. The same is true of $R(\beta|\mu)$ and $R(\beta|\mu, \alpha)$. We have that

$$R(\alpha|\mu) = R(\alpha|\mu, \beta) = bn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 \tag{123a}$$

**TABLE 7.9 Analysis of Variance for a Two-Way Classification with Interaction, with Balanced Data (all $n_{ij} = n$) (Both Parts of Table 7.8 Simplify to this When $n_{ij} = n$)**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = abn\bar{y}_{...}^2$ |
| $\alpha$ after $\mu$ | $a-1$ | $R(\alpha\|\mu) = R(\alpha\|\mu,\beta) = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$ |
| $\beta$ after $\mu$ | $b-1$ | $R(\beta\|\alpha,\mu) = R(\beta\|\mu) = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$ |
| $\gamma$ after $\mu$, $\alpha$, and $\beta$ | $(a-1)(b-1)$ | $R(\gamma\|\mu,\alpha,\beta) = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$ |
| Residual error | $ab(n-1)$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk} - \bar{y}_{ij.})^2$ |
| Total | $abn$ | $\text{SST} = \sum_{i=1}^{a}\sum_{j-1}^{b}\sum_{k=1}^{n}y_{ijk}^2$ |

and

$$R(\beta|\mu,\alpha) = R(\beta|\mu) = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2. \tag{123b}$$

This is the same as in (50).When using the statistical software package SAS for balanced data, the type I sum of squares corresponding to $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$ and the type III sum of squares corresponding to $R(\alpha|\mu,\beta)$ and $R(\beta|\mu,\alpha)$ are the same.

Table 7.9 shows the familiar analysis of variance. It is similar to Table 7.5. As was the case there, here distinction between fitting "$\alpha$ after $\mu$" and "$\alpha$ after $\beta$ and $\mu$" is no longer necessary. They are both "$\alpha$ after $\mu$" with reduction in the sum of squares $R(\alpha|\mu)$ as shown in Table 7.9. As we would expect, for the case of balanced data, the numerator of (122) also reduces to $R(\alpha|\mu)$ of (123a).When doing the analysis of variance with a hand-held calculator to minimize round-off error, we recommend the computing formulae listed below. They are,

$$R(\alpha|\mu) = \frac{1}{bn}\sum_{i=1}^{a}y_{i..}^2 - \frac{1}{abn}y_{...}^2,$$
$$R(\beta|\mu) = \frac{1}{an}\sum_{j=1}^{b}y_{.j.}^2 - \frac{1}{abn}y_{...}^2,$$

and

$$R(\gamma|\mu,\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij.}^2 - \frac{y_{...}^2}{abn} - R(\alpha|\mu) - R(\beta|\mu)$$

Furthermore,

$$\mathrm{SST_m} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn}$$

and

$$\mathrm{SSE} = \mathrm{SST_m} - R(\alpha|\mu) - R(\beta|\mu) - R(\gamma|\mu,\alpha,\beta).$$

These formulae may be shown to be algebraically equivalent to those in Table 7.9.

**Example 16 Sums of Squares Computation for Balanced Data** These data are from Chapter 14, Problem 14-2 of Montgomery Runger (2007), reproduced with the permission of John Wiley & Sons. The discussion and solution is a slight modification of that in Gruber (2014), also used with the permission of John Wiley & Sons.

An engineer suspects that the type of paint used and the drying times influences the surface finish of metal paint. He selected three drying times: 20, 25, and 30 minutes, and used two types of paint. Three parts were tested with each combination of paint and drying time. The data are as follows:

| | Drying | Time | Minutes |
|---|---|---|---|
| Paint | 20 | 25 | 30 |
| 1 | 74 | 73 | 78 |
| | 64 | 61 | 85 |
| | 50 | 44 | 92 |
| 2 | 92 | 98 | 66 |
| | 86 | 73 | 45 |
| | 68 | 88 | 85 |

*Source*: Montgomery, D.C. and G.C.

Factor A is the paint so $a = 2$. Factor B is the drying time so $b = 3$. There are three replications so that $c = 3$. Assign $i = 1$ to paint 1 and $i = 2$ to paint 2. Assign $j = 1$ to 20 minutes, $j = 2$ to 25 minutes, and $j = 3$ to 30 minutes. Then, $\mathrm{x}_{1..} = 621$ and $\mathrm{x}_{2..} = 701$. Also x.. = 1322. Then,

$$R(\alpha|\mu) = \mathrm{SSA} = \frac{621^2}{9} + \frac{701^2}{9} - \frac{1322^2}{18} = 355.556.$$

To compute $\mathrm{SSB} = R(\beta|\mu) = R(\beta|\mu,\alpha)$, observe that $\mathrm{x}_{.1.} = 434$, $\mathrm{x}_{.2.} = 437$, and $\mathrm{x}_{.3.} = 451$. Then,

$$R(\beta|\mu) = \mathrm{SSB} = \frac{434^2}{6} + \frac{437^2}{6} + \frac{451^2}{6} - \frac{1322^2}{18} = 27.4444.$$

For interaction, we need to calculate the cell sums. We have $x_{11.} = 188, x_{12.} = 178$, $x_{13.} = 255, x_{21.} = 246$, $x_{22.} = 259$, and $x_{23.} = 196$. Now,

$$\text{SSM} = \frac{188^2}{3} + \frac{178^2}{3} + \frac{255^2}{3} + \frac{246^2}{3} + \frac{259^2}{3} + \frac{196^2}{3} - \frac{1322^2}{18} = 2261.78$$

and

$$\begin{aligned} R(\gamma|\mu, \alpha, \beta) &= \text{SSI} = \text{SSM} - \text{SSA} - \text{SSB} = 2261.78 - 355.556 - 27.4444 \\ &= 1878.78 \end{aligned}$$

The total sum of squares corrected for the mean is

$$\text{TSS} = 101598 - \frac{1322^2}{18} = 4504.44$$

and the error sum of squares is

$$\begin{aligned} \text{SSE} &= \text{SST} - \text{SSA} - \text{SSB} - \text{SSI} = 4504.44 - 355.556 - 27.444 - 1878.78 \\ &= 2242.66. \end{aligned}$$

Summarizing this in the ANOVA table

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Paint | 1 | 355.556 | 355.556 | 1.90251 |
| Drying Time | 2 | 27.4444 | 13.7222 | 0.07342 |
| Interaction | 2 | 1878.78 | 939.39 | 5.02649* |
| Error | 12 | 2242.66 | 186.888 | |
| Total | 17 | 4504.44 | | |

It appears that the finish is not affected by either the paint or the drying time. However, there is a significant interaction between the choice of paint and the drying time.An R program and output is below.

```
> finish<-c(74,73,78,64,61,85,50,44,92,92,98,66,86,73,45,68,88,85)
> paint<-c("a","a","a","a","a","a","a","a","a","b","b","b","b","b",
"b","b","b","b")
> time<-c("d","e","f","d","e","f","d","e","f","d","e","f","d","e",
"f","d","e","f")
> result<-lm(finish~paint*time)
> anova(result)
Analysis of Variance Table
Response: finish
```

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
paint        1  355.56  355.56  1.9025 0.19296
time         2   27.44   13.72  0.0734 0.92962
paint:time   2 1878.78  939.39  5.0265 0.02596 *
Residuals   12 2242.67  186.89
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The SAS output is

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **PAINT** | 2 | 1 2 |
| **TIME** | 3 | 20 25 30 |

| | |
|---|---|
| **Number of Observations READ** | 18 |
| **Number of Observations Used** | 18 |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: finish**

| Source | DF | Sum of Squares | MEAN Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 5 | 2261.777778 | 452.355556 | 2.42 | 0.0973 |
| **Error** | 12 | 2242.666667 | 186.888889 | | |
| **Corrected Total** | 17 | 4504.444444 | | | |

| R-Square | Coeff Var | Root MSE | finish Mean |
|---|---|---|---|
| 0.502121 | 18.61370 | 13.67073 | 73.44444 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **paint** | 1 | 355.555556 | 355.555556 | 1.90 | 0.1930 |
| **time** | 2 | 27.444444 | 13.722222 | 0.07 | 0.9296 |
| **paint∗time** | 2 | 1878.777778 | 939.388889 | 5.03 | 0.0260 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **paint** | 1 | 355.555556 | 355.555556 | 1.90 | 0.1930 |
| **time** | 2 | 27.444444 | 13.722222 | 0.07 | 0.9296 |
| **paint∗time** | 2 | 1878.777778 | 939.388889 | 5.03 | 0.0260 |

The code used to generate the above output was

```
Data metal;
Input paint time finish;
Cards;
```

```
1 20 74
1 25 73
.................
2 25 88
2 30 85
proc glm;
class paint time;
model finish=paint time paint*time;
run;
```

## 3. INTERPRETATION OF HYPOTHESES

None of the hypotheses (97), (100), (106), (107), or (110) are particularly appealing so far as interpretability is concerned. They all involve the data themselves—not their magnitudes but the numbers of them, the values of the $n_{ij}$. For example, (100) is

$$H : \boldsymbol{\alpha}_i + \frac{1}{n_{i.}} \sum_j n_{ij}(\boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}) \text{ equal for all } i. \tag{124}$$

The corresponding hypothesis for the no-interaction case is

$$H\text{: } \boldsymbol{\alpha}_i + \frac{1}{n_{i.}} \sum_j n_{ij}\boldsymbol{\beta}_j \text{ equal for all } i \tag{125}$$

analogous to (48). These hypotheses involve the $n_{ij}$'s in two ways:

1. in terms of the weight in which, for example, the $\beta_j$'s enter the hypothesis;
2. whether some of the $\beta_j$'s enter the hypothesis at all.

Thus, for example, in (124), if $n_{ip} = 0$, $\beta_p$ will not occur in the expression containing $\alpha_i$. As a result, the pattern of the data, specifically the pattern of which $n_{ij}$ are zero and which are not, governs the form of the hypotheses tested by the $F$-statistics in Tables 7.2 and 7.7. In Table 7.2, $F(\alpha|\mu, \beta)$ and $F(\beta|\mu, \alpha)$ test, respectively, the differences between $\alpha$'s and the differences between $\beta$'s. Otherwise, all hypotheses tested by the $F$'s in Tables 7.2 and 7.7 involve data through the values of the $n_{ij}$. For $F(M)$ in both tables, the hypothesis is $H$: $E(\bar{y}) = 0$. This is

$$H\text{: } \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{n_{ij}(\mu + \alpha_i + \beta_j + \gamma_{ij})}{N} = 0.$$

This hypothesis involves weighted means of the elements of the model as they occur in $\bar{y}$. For $F(\alpha|\mu)$, the hypothesis involves the $\alpha$'s in the presence of a weighted mean of those $\beta$'s and $\gamma$'s with which the $\alpha$'s occur in the data. Likewise, in $F(\beta|\mu)$, the hypothesis involves the $\beta$'s in the presence of the weighted means of the $\alpha$'s and $\gamma$'s. For $F(\alpha|\mu, \beta)$ and $F(\beta|\mu, \alpha)$ of Table 7.7, the hypotheses involve the $n_{ij}$'s in the

somewhat complex manner shown in (106) and (107). In Table 7.7, the only case where the $n_{ij}$'s are not explicitly involved is the hypothesis (110) for $F(\gamma|\mu, \alpha, \beta)$. However, here the effect of the $n_{ij}$'s is implicit, because whether or not they are zero or non-zero, determines which functions of the $\gamma$'s make up the hypothesis.

This dependence of hypotheses on the structure of available data throws doubt on the validity of such hypotheses. Usually an experimenter wishes to test hypotheses that arise from the context of his/her work and not hypotheses that depend on the pattern of $n_{ij}$'s in his/her data. However, in general, the $F$-statistics of the analyses in Table 7.2, 7.3 and 7.7 do rely on the $n_{ij}$-values of the data. The only way that some of the hypotheses corresponding to the analysis of variance $F$-statistics might be valid would be if the $n_{ij}$'s, as they occur in the data, are in direct proportion to the occurrence of the elements if the model in the population. This is the case of proportionate subclass numbers. For this case, for example, (125) becomes

$$H\colon \alpha_i + \sum_{j=1}^{b} p_j \beta_j \text{ equal for all } i.$$

This is equivalent to

$$H\colon \alpha_i \text{ equal for all } i.$$

A feature of the hypothesis (110) tested by $F(\gamma|\mu, \alpha, \beta)$ warrants attention. It involves what hypotheses are actually being tested when testing for interaction. Consider the following measure of the extent to which the difference between the expected value of the $i$th and the $i'$th treatments, in terms of Table 7.6, when used on variety $j'$. The measure is given by

$$\begin{aligned}
\theta_{ij,i'j'} &= \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} \\
&= \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \\
&= E(\bar{y}_{ij.}) - E(\bar{y}_{i'j.}) - E(\bar{y}_{ij'.}) + E(\bar{y}_{i'j'.}) \\
&= [E(\bar{y}_{ij.}) - E(\bar{y}_{i'j.})] - [E(\bar{y}_{ij'.}) + E(\bar{y}_{i'j'.})].
\end{aligned}$$

This is just the measure of interaction discussed in Section 3d(ii) of Chapter 4. Hence, we can say that $F(\gamma|\mu, \alpha, \beta)$ test interactions. What does this really mean? It does not necessarily mean that we are testing the hypotheses that the interactions are zero. It would, if the hypotheses were $\theta_{ij,i'j'} = 0$ for sets of various values of $i, j, i'$, and $j'$. However, this is not always so. For example, in (111), part of the hypothesis is $\theta_{11,22} + \theta_{12,33} = 0$ or equivalently, from (116), $\theta_{22,33} - \theta_{21,13} = 0$. As hypotheses, these two statements are not equivalent to hypotheses of $\theta'$s being zero. This is an important fact! It means, for example, that in testing $\theta_{22,33} - \theta_{21,13} = 0$, each of $\theta_{22,33}$ and $\theta_{21,13}$ could be non-zero with the hypotheses still being true. In fact, $\theta_{22,33}$ and $\theta_{21,13}$ could both be very large but nevertheless equal, and as a result the hypothesis $H\colon \theta_{22,33} - \theta_{21,13} = 0$ would still be true. The upshot of this discussion is that for unbalanced data, $F(\gamma|\mu, \alpha, \beta)$ is not testing that interactions are zero. Some

**TABLE 7.10 Presence or Absence of Data for Discussing Connectedness (× Indicates One or More Observations; − Indicates no Observations)**

| | Level of $\beta$-Factor | | | |
|---|---|---|---|---|
| Level of $\alpha$-Factor | 1 | 2 | 3 | 4 |
| 1 | × | × | − | − |
| 2 | × | × | − | − |
| 3 | − | − | × | × |

interactions can be non-zero, although equal in magnitude, of the same or opposite sign, with the hypothesis tested by $F(\gamma|\mu, \alpha, \beta)$ still being true.

## 4. CONNECTEDNESS

Suppose available data occur as indicated in Table 7.10. If each cell that contains data has only a single observation, the normal equations are as follows.

$$\begin{matrix} \mu^\circ & \alpha_1^\circ & \alpha_2^\circ & \alpha_3^\circ & \beta_1^\circ & \beta_2^\circ & \beta_3^\circ & \beta_4^\circ \end{matrix}$$
$$\begin{bmatrix} 6 & 2 & 2 & 2 & 2 & 2 & 1 & 1 \\ 2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \beta_1^\circ \\ \beta_2^\circ \\ \beta_3^\circ \\ \beta_4^\circ \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{.1} \\ y_{.2} \\ y_{.3} \\ y_{.4} \end{bmatrix} \tag{126a}$$

Subtracting the fourth equation from the first, changes (126a) to

$$\begin{matrix} \mu^\circ & \alpha_1^\circ & \alpha_2^\circ & \alpha_3^\circ & \beta_1^\circ & \beta_2^\circ & \beta_3^\circ & \beta_4^\circ \end{matrix}$$
$$\begin{bmatrix} 4 & 2 & 2 & 0 & 2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \alpha_3^\circ \\ \beta_1^\circ \\ \beta_2^\circ \\ \beta_3^\circ \\ \beta_4^\circ \end{bmatrix} = \begin{bmatrix} y_{1.} + y_{2.} \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{.1} \\ y_{.2} \\ y_{.3} \\ y_{.4} \end{bmatrix}. \tag{126b}$$

Equations (126b) can be rewritten as two separate sets of equations. They are

$$\begin{bmatrix} 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 0 & 1 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 \\ 2 & 1 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \beta_1^\circ \\ \beta_2^\circ \end{bmatrix} = \begin{bmatrix} y_{1.} + y_{2.} \\ y_{1.} \\ y_{2.} \\ y_{.1} \\ y_{.2} \end{bmatrix} \tag{127}$$

and

$$\begin{bmatrix} 2 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_3^\circ \\ \beta_3^\circ \\ \beta_4^\circ \end{bmatrix} = \begin{bmatrix} y_{3.} \\ y_{.3} \\ y_4 \end{bmatrix}. \tag{128}$$

Thus, even though the normal equations for the data pattern of Table 7.10 are (126a), we can separate them into two sets of equations (127) and (128). Apart from $\mu$, equations (127) and (128) involve quite separate sets of parameters. The parameters in (127) are $\alpha_1, \alpha_2, \beta_1,$ and $\beta_2$. The parameters for (128) are $\alpha_3, \beta_3,$ and $\beta_4$. Furthermore, the data involved in the two sets of equations are also separate. In (127), we have $y_{11}$, $y_{12}$, $y_{21}$, and $y_{22}$. In (128), we have $y_{33}$ and $y_{34}$. This separation of the normal equations is the result of the way that certain cells of the two-way classification have data and others do not. When this separation occurs, we say that the data are *not connected*, or *disconnected*. When it does not occur, the data are *connected*. When data are disconnected, the separate sets of data corresponding to the separate sets of normal equations, such as (127) and (128) will be called *disconnected sets of data.* Thus, data in the pattern of Table 7.10 are disconnected. There are two disconnected sets of data. One consists of $y_{11}$, $y_{12}$, $y_{21}$, and $y_{22}$. The other consists of $y_{33}$ and $y_{34}$.

The underlying characteristic of disconnected data is that each of its disconnected sets of data can be analyzed separately from the other such sets. Each data set has its own normal equations that can be solved without reference to those of the other data sets. Indeed, this is the case for equations (127) and (128). Certainly, each set of normal equations contains $\mu^\circ$. However, since each group of normal equations is of less than full rank, they can all be solved with a common $\mu^\circ$, if desired. One possible choice for $\mu^\circ$ is 0.

Disconnectedness of data means not only that each of its disconnected sets of data can be analyzed separately. It also means that all the data cannot be analyzed as a single group of data. For example, as mentioned in Section 1d, in the "absorption process" for obtaining $R(\mu, \alpha, \beta)$, the matrix $\mathbf{C}^{-1}$ does not exist for disconnected data. Another reason that we cannot analyze disconnected sets of data as a single data set is due the degrees of freedom that would result if we tried it. For example, data in the pattern of Table 7.10 would give degrees of freedom for $R(\gamma|\alpha, \mu, \beta)$ as $s - a - b + 1 = 6 - 3 - 4 + 1 = 0$. For some patterns of data, this value can be

**TABLE 7.11 Pooling of Analyses of Variance of Disconnected Sets of Data in a Two-Way Classification**

| $t$th Disconnected Set of Data | | | Pooling of $d$ Disconnected Sets of Data | | |
|---|---|---|---|---|---|
| Source of Variation | d.f. | Sum of Squares | Source of Variation | d.f. | Sum of Squares |
| $\mu$ | 1 | $R_t(\mu)$ | $\mu$: for each set | d | $\sum_{t=1}^{d} R_t(\mu)$ |
| $\alpha\|\mu$ | $a_t - 1$ | $R_t(\alpha\|\mu)$ | $\alpha\|\mu$, within sets | $\sum_{t=1}^{a} a_t - d = a - d$ | $\sum_{t=1}^{d} R_t(\alpha\|\mu)$ |
| $\beta\|\mu, \alpha$ | $b_t - 1$ | $R_t(\beta\|\mu, \alpha)$ | $\beta\|\mu, \alpha$, within sets | $\sum_{t=1}^{d} b_t - d = b - d$ | $\sum_{t=1}^{d} R_t(\beta\|\mu, \alpha)$ |
| $\gamma, \mu, \alpha, \beta$ | $p_t = s_t - a_t - b_t + 1$ | $R_t(\gamma\|\mu, \alpha, \beta)$ | $\gamma\|\mu, \alpha, \beta$, within sets | $\sum_{t=1}^{d} p_t = p + d - 1$ | $\sum_{t=1}^{d} R_t(\gamma\|\mu, \alpha, \beta)$ |
| Residual | $N_t - s_t$ | $\text{SSE}_t$ | Residual, within sets | $\sum_{t=1}^{d} (N_t - s_t) = N - s$ | $\sum_{t=1}^{d} \text{SSE}_t$ |
| Total | $N_t$ | $\left(\sum y^2\right)_t$ | Total | $\sum_{t=1}^{d} N_t$ | $\sum_{t=1}^{d} \left(\sum y^2\right)_t$ |

negative. For instance, if there were no data in the (1, 1)-cell of Table 7.10 $s - a - b + 1 = 5 - 3 - 4 + 1 = -1$. This would be meaningless!

Disconnected data have to be analyzed on a within-set basis. This holds true whether there is one observation or more than one observation per filled cell. We can make the appropriate analysis (Table 7.3 or 7.8) within each disconnected set of data. Then, from these analyses, we can establish a pooled analysis. However, such pooling may be of little practical value because of the complexity of some of the hypotheses that are tested by the $F$-statistics implicit in Tables 7.3 or 7.8. Nevertheless, it is instructive to demonstrate the degrees of freedom for these analysis, as distinct from those that would be given by analyzing the complete data by taking their disconnectedness into account. We show the pooling in Table 7.11. We assume that there are $d$ sets of disconnected data and that the $i$th set has $N_i$ observations, $a_t$ rows, $b_t$ columns, and $s_t$ filled cells. The corresponding sums of squares are also subscripted by $t$. The nature of the disconnectedness ensures that

$$N = \sum_{t=1}^{d} N_t, \quad a = \sum_{t=1}^{d} a_t, \quad b = \sum_{t=1}^{d} b_t \quad \text{and} \quad s = \sum_{t=1}^{d} s_t.$$

In Table 7.11, we also write

$$p = s - a - b + 1 \quad \text{and} \quad p_t = s_t - a_t - b_t + 1,$$

with

$$p = \sum_{t=1}^{d} p_t - d + 1. \tag{129}$$

Table 7.11 is based on Table 7.8, for fitting $\alpha|\mu$ and $\beta|\mu,\alpha$. We can also construct a similar table for fitting $\beta|\mu$ and $\alpha|\mu,\beta$.

In Table 7.11, the residual sum of squares for the pooled analysis provides an estimator of $\sigma^2$ as

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{d} \mathrm{SSE}_t}{\sum_{t=1}^{d} (N_t - s_t)}.$$

We can use this estimator in tests of hypothesis. Furthermore, we may partition the first line of the pooled analysis, that for means, into two terms. Let

$$m = \text{mean of all data} = \frac{\sum_{t=1}^{d} \sqrt{N_t R_t(\mu)}}{\sum_{t=1}^{d} N_t}.$$

Then the partitioning of $\sum_{t=1}^{d} R_t(\mu)$ with $d$ degrees of freedom is

$$m^2 \sum_{t=1}^{d} N_t \quad \text{with 1 degree of freedom}$$

and

$$\sum_{t=1}^{d} R_t(\mu) - m^2 \sum_{t=1}^{d} N_t \quad \text{with } d-1 \text{ degrees of freedom.}$$

We can use the second of these two expressions divided by $(d-1)\hat{\sigma}^2$ to test the hypothesis of equality of the $E(\overline{y})$'s corresponding to the disconnected sets of data.

Table 7.12 gives an example of Table 7.11 showing degrees of freedom only for the data pattern of Table 7.10.

Disconnectedness has a great effect on estimability of functions. For example, in the case of the no-interaction model of equations (127) and (128) derived from Table 7.10, $\beta_1 - \beta_3$ is not estimable. The reason for this is that $\beta_1$ is a parameter in one

**TABLE 7.12 Degrees of Freedom in Analysis of Variance for Data Pattern of Table 7.10**

| | Degrees of freedom | | | |
|---|---|---|---|---|
| | Analyzed as disconnected data | | | |
| | 2 disconnected sets | | | |
| Source of Variation | Set I Cells 11, 12, 21, and 22 | Set II Cells 33 and 34 | Pooled Analysis | Analyzed Wrongly, As one Set of Data Ignoring Disconnectedness |
| $\mu$ | 1 | 1 | 2 | 1 |
| $\alpha \mid \mu$ | 1 | 0 | 1 | 2 |
| $\beta \mid \mu, \alpha$ | 1 | 1 | 2 | 3 |
| $\gamma \mid \mu, \alpha, \beta$ | 1 | 0 | 1 | 0 |
| Residual | $N_1 - 4$ | $N_2 - 2$ | $N - 6$ | $N - 6$ |
| Total | $N_1$ | $N_2$ | $N$ | $N$ |

disconnected set of data and $\beta_3$ in the other. In general, functions of parameters that involve parameters relating to different disconnected sets of data are not estimable. On the other hand, functions involving parameters relating to any single set of connected data, including such sets that are subsets of disconnected data, can be estimable. For the example in Table 7.10, where the data are from a no-interaction model, $\beta_2 - \beta_3$ is not estimable, but $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ are. For the interaction model, $\mu_{ij}$ is estimable for all $n_{ij} \neq 0$. However, functions of $\mu_{ij}$ that involve $\mu_{ij}$ from different disconnected sets of data are not estimable.

For connected data, the rank of $\mathbf{X}$, or equivalently of $\mathbf{X}'\mathbf{X}$, in the normal equations is $a + b - 1$ in the no-interaction case. Thus, if the data corresponding to Table 7.10 were connected, the rank of $\mathbf{X}'\mathbf{X}$ in (126) would be $3 + 4 - 1 = 6$. However, since the data are not connected, the rank is $a + b - 1 - (d - 1) = 5$ where $d$ is the number of disconnected sets of data. equations (127) and (128) illustrate this. Their ranks are $2 + 2 - 1 = 3$ and $1 + 2 - 1 = 2$, respectively, summing to 5 the rank of (126). This accounts for the relationship $p = \sum_{i=1}^{d} p_t - (d - 1)$ shown in (129) and Table 7.11.

In order for us to be able to analyze a complete data set by the methods of Table 7.3 or 7.8 the data set must be connected. Weeks and Williams (1964) discuss connectedness of data for the general $k$-way classification without interaction. They give a procedure for determining whether or not data are connected. We shall discuss this in Chapter 8. For data in a two-way classification, it simplifies to the following. Take any cell containing data—the $(p, q)$th cell, say. From that cell, move along the $p$th row (in either direction), or up or down the $q$th column until another filled cell is encountered. If by moving in this direction, all filled cells can be encountered, then the data are connected. Otherwise, they are disconnected. If data are disconnected, the process will isolate the disconnected set of data containing the original $(p, q)$th cell. Restarting the process in some cell not in that set will generate another disconnected set. Continued repetition in this manner yields all the disconnected sets.

**Example 17 Isolating Disconnected Sets of Data** In the following array of dots and ×'s, a dot represents an empty cell and an × represents a filled cell. The lines joining the ×'s isolate the disconnected sets of data.

□

More information about connectedness of experimental designs is available in Shah and Khatri (1973), Shah and Dodge (1977), Park and Shah (1995), Raghavarao and Federer (1975), Eccleston and Russell (1975), and Godolphin (2013), and the references given in these articles.

Nowadays, people are analyzing big data sets. For such data sets, it would be important to locate all of the disconnected sets of data, analyze them and pool them as we have discussed or by some other method. The ins and outs of how to do this could be an important area of research.

## 5. THE $\mu_{ij}$ MODELS

In discussing estimable functions in both the no-interaction and interaction models of Sections 1 and 2, great use was made of the fact that both $\mu + \alpha_i + \beta_j$ and $\mu + \alpha_i + \beta_j + \gamma_{ij}$ were estimable. In both cases, all of the estimable functions were linear combinations of these. In neither case were $\mu$, the $\alpha_i$, nor the $\beta_j$ individually estimable, nor the $\gamma_{ij}$ in the interaction case. For special cases of restricted models, usually with balanced data, these individual elements can become estimable (as discussed in Sections 1h and 2g) but in general, they are not. However, if we write $\mu_{ij} = \mu + \alpha_i + \beta_j$ in the no-interaction model, we can say that in each model, the basic underlying estimable function is $\mu_{ij}$ (appropriately defined for $n_{ij} \neq 0$). This fact leads to considering what may be called $\mu_{ij}$ models for $n_{ij} \neq 0$.

A $n_{ij} \neq 0$ model consists of simply writing in the interaction case

$$y_{ijk} = \mu_{ij} + e_{ijk},$$

where the $e_{ijk}$ have the same properties as before. Then, for $n_{ij} \neq 0$, $\mu_{ij}$ is estimable. Its b.l.u.e. is $\bar{y}_{ij.}$. Its variance is $v(\mu_{ij}) = \sigma^2/n_{ij}$. Any linear function of the estimable

$\mu_{ij}$'s is estimable. For example, $\mathbf{k}'\boldsymbol{\mu}$ is estimable with b.l.u.e. $\mathbf{k}'\bar{\mathbf{y}}$ and variance $\mathbf{k}'\mathbf{D}\{1/n_{ij}\}\mathbf{k}\sigma^2$. Furthermore, any hypothesis relating to linear functions of the $\mu$'s is testable. Moreover, the reduction in the sum of squares for fitting the model is

$$R(\mu) = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij.} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij.}/n_{ij}.$$

This is the same reduction in the as that in fitting any of the models containing $\gamma_{ij}$. See equations (71) and (72). Gruber (2014) uses a similar approach to the above when deriving the interaction model for the balanced case.

The simplicity of such a model is readily apparent. There is no confusion over which functions are estimable, what their b.l.u.e.'s are and what hypotheses can be tested. This results from the fact that the $\mu_{ij}$-model is always of full rank with the corresponding values of $\mathbf{X}'\mathbf{X}$ being $\mathbf{D}\{n_{ij}\}$ for $n_{ij} \neq 0$. Therefore, the normal equations are quite straightforward. They have simple solutions $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$, where $\boldsymbol{\mu}$ is the vector of $\mu$'s and $\bar{\mathbf{y}}$ the corresponding vector of observed cell means.

The $\mu_{ij}$-models have the property that the number of parameters in the model equals the number of filled cells. This gives rise to the full-rank nature of the normal equations. This is because that a model specified this way is not over-specified as it is when using the customary $\mu$, $\alpha_i$'s, and $\beta_j$'s. For example, in the no-interaction model there are, with $a$ rows and $b$ columns $1 + a + b$ parameters, but only $a + b - 1$ linearly independent means with which to try to estimate them. For the interaction model, there are with $s$ filled cells, $1 + a + b + s$ parameters but only $s$ linearly independent means. In both cases, therefore, there are more parameters in the model than there are linearly independent means in the estimation process. Hence, it is impossible to estimate every parameter individually. Therefore, the $\mu_{ij}$ model is conceptually much easier because there are exactly as many $\mu_{ij}$'s to be estimated as there are observed cell means, with a one-dimensional correspondence.

This is appropriate from the sampling viewpoint, because to the person whose data are being analyzed, the important thing is the $s$ populations corresponding to the $s$ observed sample means $\bar{y}_{ij.}$. Each of these is an estimator of the mean of the population from which the $y_{ijk}$'s are deemed to be a random sample. These populations are the factor of underlying interest. Therefore, the $\bar{y}_{ij.}$, the sample means as estimators (b.l.u.e.'s) of the population means, are the foundation of the estimation procedure. So far as estimating functions of these population means and testing hypotheses about them, it is up to the person whose data they are, presumably in consultation with a statistician, to specify in terms of the $\mu_{ij}$'s, the functions and hypotheses that are of interest to him. This, of course, is done within the context of the data and what they represent. In short, the situation is no more than estimating population means and functions of them and testing hypotheses about them. Just what functions and hypotheses we study, are determined by the contextual situation of the data. Speed (1969), and Hocking and Speed (1975) give a very complete discussion of the whole topic. Urquhart et al. (1970, 1973), in considering certain aspects of it, trace the historical development of linear models as we use them today.

As an example, the experimenter, or person whose data are being analyzed can define row effects as

$$\rho_i = \sum_{j=1}^{b} t_{ij}\mu_{ij} \quad \text{for } n_{ij} \neq 0$$

by giving to $t_{ij}$ any value he/she pleases. Then, the b.l.u.e. of $\rho_i$ is $\hat{\rho}_i = \sum_{j=1}^{b} t_{ij}\bar{y}_{ij.}$ with

$$v(\hat{\rho}_i) = \hat{\sigma}^2 \sum_{j=1}^{b} \frac{t_{ij}^2}{n_{ij}}.$$

The hypothesis $H$: all $\rho_i$ equal can be tested using

$$F = \frac{\sum_{i=1}^{a} \hat{\rho}_i^2 / v(\hat{\rho}_i) - \left[\sum_{i=1}^{a} \hat{\rho}_i^2 / v(\hat{\rho}_i)\right]^2 \Big/ \sum_{i=1}^{a} [1/v(\hat{\rho}_i)]}{(a-1)} \tag{130}$$

as given by Henderson (1968). Proof of this result is established in the same manner as equation (122).

Novel as this simplistic approach might seem, it is of essence not new. In fact, it long preceded the development of the analysis of variance itself. Urquhart et al. (1970) have outlined how Fisher's early development of analysis of variance stemmed from ideas on intra-class correlation. Establishment of models with elements $\mu, \alpha_i, \beta_j$, and so on, such as developed in this text followed the analysis of variance and did not precede it. Prior to it, there is a plentiful literature on least squares (354 titles in a bibliography dated 1877 in Urquhart et al., 1970) based essentially on the estimation of cell means. Any current or future adoption of this handling of linear models would therefore represent no new basic concept. Success in doing this does demand of today's readers, a thorough understanding of current procedures.

## 6. EXERCISES

**1** Four men and four women play a series of bridge games. At one point in their playing, their scores are as shown below.

Bridge Scores (100's)

| | Men | | | |
|---|---|---|---|---|
| Women | A | B | C | D |
| P | 8 | – | 9 | 10 |
| Q | 13 | – | – | – |
| R | – | 6 | 14 | – |
| S | 12 | 14 | 10 | 24 |

The blanks are the scores lost by the scorekeepers. Carry out an analysis of variance procedure to investigate differences by players of the same sex. In so doing, calculate the sums of squares fitting men before women and women before men, and make both analysis of variance tables.

**2** Make a rank transformation for the data of Exercise 1 and do the analysis of variance.

**3** For the data of Table 7.1:

**(a)** Set $\mu^\circ = 0, \alpha_1^\circ = 0$ and then find $\mathbf{b}^\circ$.

**(b)** Find $R(\alpha|\mu, \beta)$ and $R(\mu, \alpha, \beta)$. Compare your answers to the results obtained for the data in the text.

**4** In Table 7.1, change the observation for stove W and pan A from 6 to 12 and repeat the analysis of Table 7.2. What conclusions do you draw?

**5** In Table 7.1, change the observation for stove W and pan A from 6 to 15 and repeat the analysis of Table 7.2. What conclusions do you draw?

**6** Repeat Exercise 3 for the data of Exercises 4 and 5.

**7** Suppose that the lost observations of Table 7.1 are found to be 13 and 5 for pans A and B, respectively on stove Y and 12 for pan B on stove Z.

**(a)** Solve the normal equations for the complete set of (now balanced data) by the same procedure as used in equations (3)–(11).

**(b)** Do the analysis of Table 7.2. What conclusions do you draw?

**8** The data for this exercise are taken from Montgomery (2005) with permission from John Wiley & Sons. A golfer recently purchased new clubs in the hope of improving his game. He plays three rounds of golf at three different golf courses with the old and the new clubs. The scores are given below.

| | Course | | |
|---|---|---|---|
| Club | Ahwatukee | Karsten | Foothills |
| Old | 90 | 91 | 88 |
| | 87 | 93 | 86 |
| | 86 | 90 | 90 |
| New | 88 | 90 | 86 |
| | 87 | 91 | 85 |
| | 85 | 88 | 88 |

Perform analysis of variance to determine if

**(a)** The score is different for the old and the new clubs.

**(b)** There is a significant difference amongst the scores on the three different golf courses.

**(c)** There is significant interaction between the golf courses and the age of the clubs.

**9** For the data of Table 7.6, establish which of the following functions are estimable and find their b.l.u.e.'s.

**(a)** $\alpha_2 - \alpha_3 + \beta_1 + \gamma_{21} - \frac{1}{2}(\beta_3 + \beta_4 + \gamma_{33} + \gamma_{34})$

**(b)** $\beta_2 - \beta_1 + \frac{1}{2}(\alpha_2 - \alpha_1) + \frac{1}{2}(\gamma_{22} + \gamma_{32} - \gamma_{13} - \gamma_{33})$

**(c)** $\alpha_1 - \alpha_2 + \frac{1}{3}(\beta_1 - \beta_2) + \frac{1}{3}(\gamma_{11} - \gamma_{12})$

**(d)** $\beta_2 - \beta_3 + \frac{1}{2}(\gamma_{22} + \gamma_{32}) - \frac{1}{3}(\gamma_{13} + 2\gamma_{33})$

**(e)** $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$

**(f)** $\gamma_{11} - \gamma_{14} - \gamma_{21} + \gamma_{22} - \gamma_{32} + \gamma_{34}$

**10** Set up a linear hypothesis that $F(\beta|\mu)$ tests in Table 7.7c. Show that its numerator sum of squares is 37.8.

**11** Set up a linear hypothesis that $F(\beta|\mu, \alpha)$ tests in Table 7.7b. Show that its numerator sum of squares is 36.7857.

**12** The following is an illustration of unbalanced data used by Elston and Bush (1964).

| | Level of $\beta$-factor | | | |
|---|---|---|---|---|
| Level of $\alpha$ factor | 1 | 2 | 3 | Total |
| | | Observations | | |
| 1 | 2,4 | 3,5 | 2,3 | 19 |
| 2 | 5,7 | – | 3,1 | 16 |
| Total | 18 | 8 | 9 | 35 |

**(a)** Calculate Table 7.8 for these data.

**(b)** An analysis of variance given for these data shows the following sums of squares.

| | |
|---|---|
| A | 3.125 |
| B | 12.208 |
| Interaction | 6.125 |
| Error | 8.500 |

**(c)** Show that the sum of squares designated A is $R(\alpha|\mu, \beta)$ and that denoted by B is $R(\beta|\mu, \alpha)$.

**(d)** Write down hypotheses tested by the $F$-statistics available from your calculations and verify their numerator sum of squares.

**13** **(a)** Calculate analyses of variance for the following data. Which of the effects are statistically significant?

| | $\beta$-factor | | |
|---|---|---|---|
| $\alpha$-factor | Level 1 | Level 2 | Level 3 |
| | | Observations | |
| Level 1 | 13,9,8,14 | 9,7 | – |
| Level 2 | 1,5,6 | 13,11 | 6,12,7,11 |

**(b)** Establish the hypothesis tested by $F(\gamma|\alpha, \beta, \mu)$.

**14** Suppose a two-way classification has only two rows and two columns.

**(a)** Prove that

**(i)** $R(\boldsymbol{\alpha}|\boldsymbol{\mu}) = \dfrac{n_{1.}n_{2.}(\bar{y}_{1..} - \bar{y}_{2..})^2}{n_{..}}$,

**(ii)** $R(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\alpha}) = \dfrac{(y_{.1.} - n_{11}\bar{y}_{1..} - n_{21}\bar{y}_{2..})^2}{(n_{11}n_{12}/n_{1.} + n_{21}n_{22}/n_{2.})}$ and

**(iii)** $R(\boldsymbol{\gamma}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \dfrac{(\bar{y}_{11.} - \bar{y}_{12.} - \bar{y}_{21.} + \bar{y}_{22.})^2}{(1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22})}$

**(b)** Write down the analogous expressions for $R(\boldsymbol{\beta}|\boldsymbol{\mu})$ and $R(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\beta})$

**(c)** Using the expressions in (a) and (b), calculate the analysis of variance tables for the data below. Which factors, if any, are statistically significant?

| | $\boldsymbol{\beta}$-factor | |
|---|---|---|
| $\boldsymbol{\alpha}$-factor | Level 1 | Level 2 |
| Level 1 | 9,10,14 | 2,4,2,3,4 |
| Level 2 | 63 | 10,12,15,14,15,18 |

**(d)** Suppose there was a typo and at $\boldsymbol{\alpha}$ Level 2 and $\boldsymbol{\beta}$ Level 1, we should have 6, 3 in place of 63. Do the analysis of variance over again and compare your results to those in (c). Would there be any difference in your conclusions?

**15** For the data of Table 7.6, find $\mathbf{u}$ and $\mathbf{T}$ and verify the value of $R(\alpha|\mu, \beta)$ in Table 7.7.

**16** For the data of Example 7, find 95% simultaneous confidence intervals on the mean number of traffic fatalities for adjacent speed limits using both Bonferonni and Scheffe's method.

**17** Show that the data in Table 7.1, 7.6, Exercises 1, 13, and 14 are connected.

**18** Suppose that data occur in the following cells of a two-way classification.
(1,1), (2,3), (2,6), (3,4), (3,7), (4,1), (4,5), (5,2), (5,4), (5,7), (6,5), and (7,6).

**(a)** Establish which sets of data are connected.

**(b)** Make a table similar to Table 7.12 that includes the degrees of freedom for an analysis of variance for each set of data, a pooled analysis and the degrees of freedom you would get for incorrectly analyzing the data by ignoring the disconnectedness.

**(c)** Give examples of estimable functions and non-estimable functions for a no-interaction model.

**19** Define

$$\mathbf{n}'_a = \begin{bmatrix} n_{1.} & \cdots & n_{a.} \end{bmatrix}, \mathrm{m}'_a = \begin{bmatrix} n_{1b} & \cdots & n_{ab} \end{bmatrix},$$

$$\mathbf{n}_\beta = \begin{bmatrix} n_{.1} \\ \cdot \\ \cdot \\ \cdot \\ n_{.b-1} \end{bmatrix}, \quad \mathbf{y}_\beta = \begin{bmatrix} y_{.1} \\ \cdot \\ \cdot \\ \cdot \\ y_{.b-1} \end{bmatrix} \quad \text{and} \quad \mathbf{D}_\beta = \begin{bmatrix} n_{.1} & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & n_{.b-1} \end{bmatrix}.$$

Using these definitions, those given in (19) and $n_{..}, n_{.b}, y_{..}$, and $y_{.b}$:

**(a)** Rewrite the normal equations (12).

**(b)** Show that $\mathbf{D}_\beta - \mathbf{M}'\mathbf{D}_\alpha\mathbf{M} = \mathbf{C}$.

**(c)** Using the result in (b), the formula for the inverse of a partitioned matrix given in Section 7 of Chapter 1 and the method of finding a generalized inverse given in Chapter 1 show that (21) is a generalized inverse of $\mathbf{X}'\mathbf{X}$.

**20** **(a)** Show that the inverse of

$$\mathbf{A} = x\mathbf{J} + \mathbf{D}\{y_i\}$$

is

$$\mathbf{A}^{-1} = \{a^{ij}\} \quad \text{for } i, j = 1, 2, \ldots, n$$

with

$$a^{ii} = \frac{1}{y_i} - \frac{x}{y_i^2\left(1 + x\sum_{i=1}^{n} 1/y_i\right)}$$

and

$$a^{ij} = \frac{-x}{y_i y_j\left(1 + x\sum_{i=1}^{n} 1/y_i\right)} \quad \text{for } i \neq j.$$

*Hint:* Use Woodbury's (1950) identity.

$$(\mathbf{A}+\mathbf{BCB}')^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}^{-1}$$

(See p. 37 of Gruber (2014), p. 51 of Golub and Van Loan (1996))

**(b)** With $\bar{x}_{i\cdot}=\sum_{j=1}^{n_i}x_{ij}/n_i$ and $\bar{x}_{\cdot\cdot}=\sum_{i\neq i'}^{a}\sum_{i\neq 1}^{n_i}x_{ij}/n_{\cdot}$, show that

$$\sum_{i=2}^{a}(\bar{x}_{1\cdot}-\bar{x}_{i\cdot})^2\left(n_i-\frac{n_i^2}{n_{\cdot}}\right)-\sum_{i\neq i'}^{a}\sum_{i\neq 1}^{a}(\bar{x}_{1\cdot}-\bar{x}_{i\cdot})(\bar{x}_{1\cdot}-\bar{x}_{i'\cdot})\frac{n_i n_{i'}}{n_{\cdot}}=\sum_{i=1}^{a}n_i(x_i-\bar{x}_{\cdot\cdot})^2.$$

**(c)** Consider the one-way classification model for the test of hypothesis $H$: all $\alpha'$s equal. Show that the numerator of the $F$-statistic is

$$Q=\sum_{i=1}^{a}n_i(\bar{y}_{i\cdot}^2-\bar{y}_{\cdot\cdot})^2.$$

**(d)** For the hypothesis

$$H\text{: equality of } \beta_j+\sum_{i=1}^{a}\frac{n_{ij}\alpha_i}{n_{\cdot j}} \quad \text{for all } j$$

in the no-interaction two-way classification model, show that the $F$-statistic reduces to

$$F(\beta|\mu)=\frac{\left(\sum_{j=1}^{b}n_{\cdot j}\bar{y}_{\cdot j}^2-n_{\cdot\cdot}\bar{y}_{\cdot\cdot}^2\right)}{(b-1)\hat{\sigma}^2}.$$

**21** When $n_{ij}=1$ for all $i$ and $j$, show that the method of solving the normal equations for the no-interaction model that uses equations (14) and (15) leads to solutions

$$\alpha_i^{\circ}=\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}+\bar{y}_{\cdot b} \quad \text{for all } i$$

and

$$\beta_j^{\circ}=\bar{y}_{\cdot j}-\bar{y}_{\cdot b} \quad \text{for } j=1,2,\ldots,b-1$$

with $\mu^{\circ}=0=\beta_b^{\circ}$.

**22** Show that when $n_{ij} = 1$, the equation $\mathbf{C}\boldsymbol{\beta}^\circ = \mathbf{r}$ in (16) has

$$\mathbf{C} = a\mathbf{I} - \frac{a}{b}\mathbf{J} \quad \text{and} \quad \mathbf{r} = a(\bar{y}_{\boldsymbol{\beta}} - \bar{y}_{..}\mathbf{1}_{b-1}).$$

Show that hence

$$\boldsymbol{\beta}^\circ = \mathbf{C}^{-1}\mathbf{r} = \bar{y}_{\boldsymbol{\beta}} - \bar{y}_{.b}1_{b-1}.$$

You may have already shown this if you did Exercise 21. Using the above information show that

$$R(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\mu}) = \boldsymbol{\beta}^{\circ\prime}\mathbf{r} = a\sum_{j=1}^{b}\bar{y}_{.j}^2 - ab\bar{y}_{..}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$

of Table 7.5. As a result, it follows that when $n_{ij} = 1$ for all $i$ and $j$, Tables 7.3b and 7.3c simplify to Table 7.5. (*Note:* All matrices and vectors are of order $b - 1$. The matrix $\mathbf{J}$ has unity for all of its elements. Furthermore, $\bar{\mathbf{y}}'_{\boldsymbol{\beta}} = \begin{bmatrix}\bar{y}_{.1} & \cdots & \bar{y}_{.b-1}\end{bmatrix}$.)

**23** Show that when $n_{ij} = n$ for all $i$ and $j$, equation (122) reduces to $\sum_{i=1}^{a} bn(\bar{y}_{i..} - \bar{y}_{...})^2/(a-1)\hat{\sigma}^2$.

# 8

# SOME OTHER ANALYSES

Chapter 6 and 7 illustrate applications of the general results of Chapter 5 (models of full rank) to specific models that often arise in the analysis of unbalanced data. We discuss three additional topics in the present chapter. They include

1. The analysis of large-scale survey-type data;
2. The analysis of covariance;
3. Some approximate analyses for unbalanced data.

There is no attempt at completeness in discussion of these topics. They are included to refer the reader to some of the other analyses available in the literature. The intent is to provide a connecting link between those expositions and procedures developed earlier in the book.

## 1. LARGE-SCALE SURVEY-TYPE DATA

Behavioral scientists from many different disciplines often conduct surveys. These surveys often involve the personal interviewing of individuals, heads of households and others. The data collected from such surveys are frequently very extensive. Many people may have been interviewed. Each of them may have been asked lots of questions. Consequently, the resulting data consist of observations of numerous

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

**TABLE 8.1 Some of the Factors Available on the Description of a Household in the Bureau of Labor Statistics Survey of Customer Expenditures, 2014 (http://www.bls.gov/cex/)**

| Factor | Number of Levels |
|---|---|
| 1. Consumer unit | 3 |
| 2. Income | 10 |
| 3. Education of "reference person" | 3 |
| 4. Race of "reference person" | 2 |
| 5. Hispanic Latino origin or not | 2 |
| 6. Family status (married, single, age of children, etc.) | 7 |
| 7. Occupation of "reference person" | 9 |
| 8. Geographical region | 4 |
| Total number of levels in eight factors | 40 |

variables and factors for a large number of people. We now discuss some of the problems of analyzing such data by the procedures of Chapter 5. The following example serves as an illustration.

### a. Example

The Bureau of Labor Statistics Survey of Consumer Expenditures 2014 provides an opportunity for studying patterns of family investment that include, but are not limited to, expenditures on equities, durables, and human components of the nature of medical expenses, education, and so on. The survey gathered data on many characteristics of each household interviewed. Table 8.1 shows some of these characteristics, coded as factors with different numbers of levels. The basic survey, based on a stratified sampling plan, included some 127,006 thousand-consumer units. One of the many questions of interest one might ask is, "To what extent is expenditure on durables affected by the factors listed in Table 8.1?" One way of attempting to answer this question might be fitting a linear model to the variable "expenditure of durables."

### b. Fitting a Linear Model

Of course, before attempting to fit a linear model involving as many as eight factors like those of Table 8.1 to data of the kind just described, the researcher should perform a careful preliminary examination. The examination could include various frequency counts and plots of the data. Assume that such an examination has been made and the goal is to fit a linear model along the lines of Chapter 5–7 to take account of the factors shown in Table 8.1. We shall now discuss some of the difficulties involved in trying to fit such a model.

Suppose we wish to fit this model based on a sample of, say, 5000. Let us consider what problems we might run into.

A model that would have main effects for each of the eight factors of Table 8.1 could also include all possible interactions among these factors. These would include

664 first-order interactions, 5962 second-order interactions, and 90,720 ($= 3 \times 10 \times 3 \times 2 \times 2 \times 7 \times 9 \times 4$) seventh-order interactions (interactions between a level of each of eight factors—see Chapter 4 Section 3d(ii)).

Two immediately apparent questions are

1. What is the meaning of a high-order interaction such as one of order 7?
2. How can we handle large numbers of interactions of this nature?

The answer to the second of these questions allows us to avoid, in large measure, answering the first. We can handle all the interactions only if the data consist of at least one observation in every sub-most cell of the data. In this case, there are 90,720 sub-most cells. Since there are only 5000 observations in the data, all of the interactions cannot be considered. In general, this state of affairs is likely to prevail with multi-factor survey data. This is because the number of sub-most cells in the model equals the product of the number of levels of all of the factors. Furthermore, having data in every sub-most cell requires having data in certain cells that are either empty by definition or, by the nature of the factors, are almost certain to be empty, even in the population. Even when all cells are filled, and the data could be analyzed using a model that included all interactions, the interpretation of high order interactions is usually difficult. It is rarely feasible to consider all of the interactions. For example, can we give a reasonable description in terms of the source of our data of what we mean by an eighth-order interaction? We doubt it. Indeed, it is probably fair to say that we would have difficulty in meaningfully describing interactions of order greater than 1, certainly of order greater than 2. First-order interactions can be described and understood reasonably well (see Section 3d of Chapter 4). However, interpretation of higher order interactions can present some difficulty. Therefore, we suffer no great loss if the sparseness of data prevents including the higher order interactions in our model. Fortunately, whereas survey data seldom enable all interactions to be included in a model, they often contain sufficient observations to consider first-order interactions. The first-order interactions are the ones that we can interpret most readily. This is the case for the data of 5000 observations in our hypothetical survey. There are enough observations to consider the 40 main effects of Table 8.1, together with the corresponding 664 first-order interactions. However, there are not enough observations to consider the 5962 second-order interactions.

Even when data are sufficient in number to consider first-order interactions, we may not want to include all of them in the model. For example, for the eight factors of Table 8.1, there are 36 different kinds of first-order interactions ($\frac{1}{2}n(n-1)$ kinds for $n$ factors). The choice of which interactions to include in the model is always that of the person whose data are being analyzed. He/she should know his/her data well enough to decide which interactions should be considered and which should not. The choice may not be easy even with just first-order interactions. Moreover, we will see that multifactor models without any interactions are difficult enough to interpret. Having interactions only further compounds the difficulty.

### c. Main-Effects-Only Models

We can avoid the quandary of which interactions to include in a model by omitting them all. The model then involves just the main effects—effects for each of the 40 levels of the eight factors in Table 8.1. Clearly, such a model is a great deal easier conceptually than one involving numerous interactions. The choice of which interactions to include in a model may be a matter of question. However, even though the main-effects-only model appears easier, it still has some difficulties. The first is an extension of the duality apparent in Tables 7.7b and 7.2c where only two factors are present. There, for the two-way classification, we can consider reductions in the sum of squares two ways, namely $R(\alpha|\mu)$ and $R(\beta|\alpha,\mu)$ or $R(\beta|\mu)$ and $R(\alpha|\mu,\beta)$. There are sequences for fitting the mean effects. They are either $\alpha$ then $\beta$ or $\beta$ then $\alpha$. However, for the eight factors in Table 8.1 there are $8! = 40{,}320$ sequences for fitting the main effects. The choice of which sequence to use in the two-way classification of Chapter 7 may be immaterial. There are only two sequences. It is relatively easy to look at both. However, with 40,320 sequences in the eight-way classification, it is essential to decide which few of them to consider. This is a decision that rests with the person whose data are being analyzed. Again, it is a decision that is often not easy to make. An $n$-way classification has $n!$ sequences for fitting the main effects of the $n$ factors. Table 8.2 shows $3! = 6$ sets of reductions in sums of squares that could be calculated for a three-way classification.

Searle (1971a) gives a similar discussion of the difficulties involved in linear model fitting for the Bureau of Labor Statistics Survey of Customer Expenditures, 1960–1961, using the results of an analysis by Brown (1968).

Reductions in the sums of squares such as are shown in Table 8.2 are sometimes said to "add up." They add up to $\text{SST} = \mathbf{y}'\mathbf{y}$, the total uncorrected squares of the observations. Of course, often the $R(\mu)$ term is not shown in the body of the table. Instead, it is subtracted from SST to have the other reductions in sums of squares add up to $\text{SST}_m = \text{SST} - R(\mu) = \sum y - N\bar{y}^2$. The $F$-statistics that are implicit in any of the sets of reductions of sums of squares illustrated in Table 8.2 can be used in either of two ways as they are in Chapter 7 for the two-way classification.

There, as discussed in Section 1e(vi) of Chapter 7, they are used for testing the effectiveness—in terms of explaining variation in $y$—of having certain main effect factors in the model. However, just as in Table 7.2, there are two possible ways of

**TABLE 8.2 Sets of Reductions in Sums of Squares for a Three-Way Classification, Main-Effects-Only Model, with Main Effects $\alpha$, $\beta$, and $\gamma$**

| | | | | | |
|---|---|---|---|---|---|
| $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ | $R(\mu)$ |
| $R(\alpha\|\mu)$ | $R(\alpha\|\mu)$ | $R(\beta\|\mu)$ | $R(\beta\|\mu)$ | $R(\gamma\|\mu)$ | $R(\gamma\|\mu)$ |
| $R(\beta\|\mu,\alpha)$ | $R(\gamma\|\mu,\alpha)$ | $R(\alpha\|\mu,\beta)$ | $R(\gamma\|\mu,\beta)$ | $R(\alpha\|\mu,\gamma)$ | $R(\beta\|\mu,\gamma)$ |
| $R(\gamma\|\mu,\alpha,\beta)$ | $R(\beta\|\mu,\alpha,\gamma)$ | $R(\gamma\|\mu,\alpha,\beta)$ | $R(\alpha\|\mu,\beta,\gamma)$ | $R(\beta\|\mu,\alpha,\gamma)$ | $R(\alpha\|\mu,\beta,\gamma)$ |
| SSE[a] | SSE | SSE | SSE | SSE | SSE |
| SST[1] | SST | SST | SST | SST | SST |

[a] $\text{SSE} = \mathbf{y}'\mathbf{y} - R(\mu,\alpha,\beta,\gamma)$ and $\text{SST} = \mathbf{y}'\mathbf{y} = \sum y^2$.

testing the exploratory power of having $\alpha$ in the model ($\alpha$ before $\beta$ and $\alpha$ after $\beta$), so in Table 8.2, there are, for the three-way classification, four ways of testing the effectiveness of $\alpha$. They are based on $R(\alpha|\mu), R(\alpha|\mu,\beta), R(\alpha|\mu,\gamma)$, and $R(\alpha|\mu,\beta,\gamma)$. For the $n$-way classification, there are $2^{n-1}$ ways of testing the effectiveness of a factor in this manner. For the eight-way classification of Table 8.1, this would be $2^7 = 128$. This is a direct outcome of there being $n!$ sequences in which $n$ main effect factors can be fitted, that is, $n!$ sets of reductions in sums of squares of the nature illustrated in Table 8.2. The tests of the exploratory power of having any particular main effect in model therefore depend, very naturally, on the sequence chosen for fitting the main effects.

The $F$-statistics can also be used as in Section 1g of Chapter 7, for testing hypotheses about the elements of a main-effects-only model. Here, however, just as in Section 1g of Chapter 7, the only hypotheses that relate to these elements in a clear and simple fashion are those based on fitting one factor after all of the others. If the statistical software package SAS is used, the $F$-tests would be based on the type III sum of squares. The hypothesis tested is that the effects of all levels of that factor are equal. For example, in Table 8.2, the hypothesis tested by $F(\alpha|\mu,\beta,\gamma)$ based on $R(\alpha|\mu,\beta,\gamma)$ is $H$: $\alpha$'s all equal. Similarly, $F(\beta|\mu,\alpha,\gamma)$ tests $H$: $\beta$'s all equal. This holds true in general. The statistic $F(\alpha|\,\mu,\beta,\gamma,\delta,\ldots,\theta)$ tests $H$: $\alpha$'s all equal where $\beta,\gamma,\delta,\ldots,\theta$ represents all the other main effects of a model. The other $F$-statistics that can be calculated provide tests of hypothesis that involve a complex mixture of the effects of the model, just as $R(\beta|\mu)$ tests the hypothesis of (48) given in Section 1g of Chapter 7. For example, $F(\alpha|\mu,\beta)$ from Table 8.2 will test a hypothesis that involves $\beta$'s and $\gamma$'s as well as $\alpha$'s.

We have just highlighted the difficulties that are involved in testing hypotheses by means of reductions in sums of squares that add up. They include

1. the choice of sequence for fitting the factors; and
2. the complex nature of the hypothesis tested by the $F$-statistics, other than $F(\alpha|\,\mu,\beta,\gamma,\delta,\ldots,\theta)$ (the $F$-statistic based on the type III sum of squares).

However, this in no way affects the use of the general formula

$$F(H) = \frac{(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^\circ - \mathbf{m})}{s\hat{\sigma}^2}$$

for testing any testable hypothesis $\mathbf{K}'\mathbf{b} = \mathbf{m}$ (see equation (71) of Chapter 5). The general formula above is as applicable to situations like that of Table 8.1 as it is to anything discussed in Chapters 5, 6, and 7. As always, of course, one must ascertain the estimability of $\mathbf{K}'\mathbf{b}$. However, within the confines of estimability, we can always use $F(H)$. Its use is not necessarily related to any of the sums of squares that add up to SST.

### d. Stepwise Fitting

When using multiple regressions there may, on occasion, be serious doubt about which $x$-variates from a large available set of $x$'s should be used in the regression model. This difficulty has led to the development of several procedures for letting the data select a "good" set of $x$-variates, good in the sense of accounting for the variance in $y$ in some manner. The main difference in the various procedures is in the criterion for selecting a good set. For example, one procedure fits one $x$-variate, then includes another, and so on.

One criterion is to choose an $x$-variate from one not already chosen, which leads to the greatest reduction in the residual sum of squares. Another adds and deletes variables according to their level of significance. There is a huge literature on different variable selection methods. Some good references include Draper and Smith (1998), Draper (2002), Smith (1988), and La Motte and Hocking (1970). We do not give details of these selection procedures here. We simply point out their application to multi-factor models. Instead of applying any one of these procedures to single $x$-variates, it can be applied to sets of dummy (0, 1) variables corresponding to each factor in the model. Then, rather than having to decide, a priori, in which sequence the factors should be fitted, we could use what might be called "stepwise fitting of factors." This would determine, from the data, a sequential fitting of the factors, which in some sense, ranked the factors in decreasing order of importance insofar as accounting for variation in y, is concerned.

In this way, for example, rather than our selecting one of the sequences implicit in Table 8.2, the data would select one for us. As a result of the stepwise regression technique, the basis of selection would be using reduction in sums of squares $R(\ )$ terms, as indicators of the extent to which different models account for the variation in $y$. Some references on dummy variables in stepwise regression include Cohen (1991), Brorsson, Ilver and Rydgren (1988), and Mannheim and Cohen (1978).

### e. Connectedness

It may sometimes be taken for granted that the difference between the effects of every pair of levels of the same factor is estimable in a main-effects-only model. Indeed, this is often so, but it is not universally the case. Sufficient conditions for such differences to be estimable are those set out by Weeks and Williams (1964) for data to be connected. Suppose there are $p$ factors (and no interactions) in the model and we denote the levels of the factors for an observation by the vector,

$$\mathbf{i}' = [i_1 \quad i_2 \quad \ldots \quad i_p]$$

Then two such vectors are defined as being *nearly identical* if they are equal in all except one element. The data sets in which the **i**-vector of each observation is nearly identical to at least one other observation form connected sets of data. Weeks and Williams (1964) give a procedure for establishing such sets. Their procedure is an extension of that given in Section 4 of Chapter 7 for the two-factor model.

As Weeks and Williams (1964) point out in their errata (1965), their conditions for data to be connected are sufficient but not necessary. Data can be connected (in the sense of intra-factor differences between main effects being estimable) without being nearly identical in the manner described. Fractional factorial experiments are a case in point. For example, suppose for the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$$

with $i, j$, and $k = 1, 2$, we have the data $y_{112}, y_{211}, y_{121}$, and $y_{222}$. No pair of these four observations is nearly identical in the manner just described. However,

$$E\frac{1}{2}(y_{112} - y_{211} + y_{121} - y_{222}) = \alpha_1 - \alpha_2.$$

Similarly, $\beta_1 - \beta_2$ and $\gamma_1 - \gamma_2$ are also estimable. Thus, all intra-factor differences between the main effects are estimable. This exemplifies why the general problem of finding necessary conditions for main effect differences to be estimable is difficult.

More recent work on the relationships between estimability and connectivity may be found in Godolphin (2013) and the references therein.

### f. The $\mu_{ij}$-models

What has been said about the difficulties of using a main-effects-only model for analyzing large-scale survey-type data applies even more to the analysis of such data using models that include interactions. The sequences in which the factors can be fitted, using reductions in sums of squares that add up to SST are then more numerous. The hypotheses tested by the resulting $F$-statistics are more complicated (e.g., see Section 2f of Chapter 7). The problem of connectedness in terms of the definition given in Section 4 of Chapter 7 is even more acute. The example in Table 8.1 illustrates this. There we have 90,720 cells in the data. That means that we can describe a household in the survey, 90,720 ways using the eight factors of Table 8.1. Yet the sample size is only 5000. Such data will almost assuredly not be connected.

In view of these difficulties with models that include interactions, the main-effects-only models appear more feasible, despite their own difficulties discussed in Section 1c of this chapter. The main-effects-only models have one further problem, that of complete neglect of interactions. This may be a very serious omission in practice! In situations involving many factors, as is the case in Table 8.1, one frequently feels that interactions between the factors do, most assuredly exist. Assuming that this is so, it would not be very appropriate to ignore them and proceed to make an analysis as if interactions did not exist. One way out of this predicament is to use the $\mu_{ij}$-models discussed in Section 5 of Chapter 7.

In this, we look at the means of the sub-most cells of the data. By "sub-most" cells, we mean those cells of the data defined by one level of each of the factors. In the two-way classification of Chapter 7 a sub-most cell is the cell defined by a row and a column. In the eight-way classification of Table 8.1, a sub-most cell is that defined by one level (kind) of, consumer unit, one level of income, one level of

education of the reference person, and so on. The total number of possible sub-most cells is the product of the number of levels in the classes—90,720 in Table 8.1. The number of sub-most cells in the data is the number of sub-most cells that have data in them. Call this number $s$. Then, no matter how many factors there are or how many levels each has, the mean of the observations in each sub-most cell is the b.l.u.e. of the population mean for that cell. Thus, if $\bar{y}_r$ is the mean of the $n_r$ observations in the $r$th sub-most cell, for $r = 1, 2, \ldots, s$, then $\bar{y}_r$ is the b.l.u.e. of $\mu_r$, the population mean of that cell. Furthermore, the b.l.u.e. of any linear function $\sum_{r=1}^{s} k_r \mu_r$ is $\sum_{r=1}^{s} k_r \bar{y}_r$ with variance $\sigma^2 \sum_{r=1}^{s} k_r^2 / n_r$. Moreover, any hypothesis concerning a linear function of the $\mu_r$'s is testable. Thus,

$$H\colon \sum_{r=1}^{s} k_r \mu_r = m \tag{1}$$

can be tested by comparing

$$F(H) = \frac{\left( \sum_{r=1}^{s} k_r \bar{y}_r - m \right)^2}{\hat{\sigma}^2 \sum_{r=1}^{s} k_r^2 / n_r} \tag{2}$$

against the value of the $F$-distribution with 1 and $(n - s)$ degrees of freedom for a given level of significance, for example, $\alpha = .05$. The estimator of $\sigma^2$ in this expression is the simple within sub-most cell mean square, namely,

$$\hat{\sigma}^2 = \frac{\sum_{r=1}^{s} \sum_{i=1}^{n_r} (y_{ri} - \bar{y}_r)^2}{(n. - s)}. \tag{3}$$

The numerator in (3) is, of course, identical to the SSE that would be derived by fitting a model that had in it all possible interactions.

The statistic $F(H)$ of (2) provides a means of testing the hypothesis about any linear function of the population sub-most cell means. Just what hypotheses get to be tested is the prerogative of the person whose data they are. All he or she need do is formulate his/her hypotheses of interest in terms of the sub-most cell means. Whilst this may be no easy task in many cases, at least is not complicated by the confusions of estimability and interactions. Furthermore, hypotheses about sub-most cell population means can be tested simultaneously by and extension of the standard results for testing $\mathbf{K}'\mathbf{b} = \mathbf{m}$ in Chapters 3 and 5. Thus, if $\boldsymbol{\mu}$ is the vector of sub-most cell populations and $\bar{\mathbf{y}}$ is the corresponding vector of observed means, then we can test

$$H\colon \mathbf{K}'\boldsymbol{\mu} = \mathbf{m}, \tag{4}$$

consisting of $s$ LIN functions $\mathbf{K}'\boldsymbol{\mu}$, by using

$$F(H) = \frac{(\mathbf{K}'\bar{\mathbf{y}} - \mathbf{m})'\left[\mathbf{K}'\mathbf{D}\left\{\frac{1}{n_r}\right\}\mathbf{K}\right]^{-1}(\mathbf{K}'\bar{\mathbf{y}} - \mathbf{m})}{s\hat{\sigma}^2}, \tag{5}$$

where $\mathbf{D}\{\frac{1}{n_r}\}$ is the diagonal matrix of the reciprocals of the number of observations in the sub-most cells containing data.

Repeated use of (2) and/or (5) does not provide tests whose $F$-statistics have numerator sums of squares that are independent, as is the case when using sums of squares that "add up," in the manner of Table 8.2. However, as we have seen, hypotheses tested by use of the latter do not involve simple functions of the parameters of the model. In contrast, the hypotheses in (1) and (4) which are tested by means of (2) and (5) are in terms of straightforward linear functions of sub-most cell population means. Further discussion of these procedures can be found in Speed (1969) and Urquhart et al. (1970).

## 2. COVARIANCE

We will now combine ideas from Chapters 3, 4, 5, 6, and 7 to formulate linear models where some of the elements of the $\mathbf{X}$ matrix are observed $x$'s and others are dummy (0, 1) variables. Such models might arise when we wish to compare different treatments, say the amount of weight loss on five different reducing diets. We need to take into account the initial weight of the subjects. These would be observed $x$'s. The diets could be specified using dummy variables.

In Chapter 3, the elements of the $\mathbf{X}$ matrix in the equation $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ are observed values of the $x$'s corresponding to the vector of observation $\mathbf{y}$. In Chapter 4, we saw how we can use the same equation for linear models involving factors and interaction by using dummy variables that take the values 0 and 1 for the $x$'s. Chapter 5 gives the general theory, and Chapters 6 and 7 give examples of it. We now consider the case where some of the elements of $\mathbf{X}$ are observed $x$'s and others are dummy (0, 1) variables. Such a situation represents a combining, into one model, of both regression and linear models involving factors and interactions. We generally refer to such a model as a covariance analysis. The basic analysis is that of the factors-and-interaction part of the model suitably amended by the presence of the $x$ variates—the covariables of the analysis.

General treatment of the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is given for $\mathbf{X}$ of full-column rank in Chapter 3 and for $\mathbf{X}$ not of full-column rank in Chapter 5. These two chapters cover regression and what we may call the factors-and-interactions models. Since $\mathbf{X}$ being of full-column rank is just a special case of $\mathbf{X}$ not being of full-column rank, the procedures of Chapter 5 apply in general to all kinds of $\mathbf{X}$ matrices. In particular, they are applicable to the analysis of covariance. Conceptually, there is no distinction between the analysis of covariance and what we have already considered. The sole difference is in the form of the elements of $\mathbf{X}$. In regression (Chapter 3),

the elements of $\mathbf{X}$ (apart from the column 1 corresponding to $\mu$) are observed $x$'s. In factors and interaction models (Chapters 5, 6, and 7), the elements of $\mathbf{X}$ are 0 or 1 corresponding to dummy variables. In analysis of covariance, some of the elements of $\mathbf{X}$ are dummy variables 0's and 1's and some are observed values of $x$ variables. Thus, conceptually, nothing is new in the analysis of covariance. It involves fitting a model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ where some elements of $\mathbf{b}$ are effects corresponding to the levels of factors and interactions, in the manner of Chapter 5–7 and some are regression-style coefficients of $x$-variates, in the manner of Chapter 3. Within this context, the procedures for solving normal equations, establishing estimable functions and their b.l.u.e's, testing hypotheses and calculating reductions in sums of squares all follow the same pattern established in Chapter 5 and summarized at the beginning of Chapter 6. No additional concepts are involved. Furthermore, the "recipes" for covariance analysis for balanced data that are to be found in many texts (e.g., Federer (1955 Chapter XVI), Steel and Torrie (1960, Chapter 15), Rao (1973, Section 4h)), and Montgomery (2005, Section 15-3)) are just the consequence of simplifying the general results for unbalanced data.

## a. A General Formulation

**(*i*) *The Model.*** We will distinguish between the two kinds of parameters that occur in $\mathbf{b}$ when using the model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$. We partition $\mathbf{b}$ into two parts. They are

1. The vector $\mathbf{a}$ for the general mean $\mu$ and the effects corresponding to levels of factors and their interactions,
   and
2. The vector $\mathbf{b}$ for the regression-style coefficients of the covariates.

The corresponding incidence matrices will be $\mathbf{X}$ for the dummy (0, 1) variables and $\mathbf{Z}$ for the values of the covariates. We write the model as

$$\mathbf{y} = \mathbf{Xa} + \mathbf{Zb} + \mathbf{e} \tag{6}$$

where $\mathbf{e} = \mathbf{y} - E(\mathbf{y})$ with $E(\mathbf{e}) = 0$ and $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$ in the customary manner. In this formulation, $\mathbf{X}$ does not necessarily have full rank. However, we will assume that $\mathbf{Z}$ does have full rank. This will usually be the case. Thus, $\mathbf{X}'\mathbf{X}$ may not have an inverse while $(\mathbf{Z}'\mathbf{Z})^{-1}$ usually exists. Furthermore, we make the customary and realistic assumption that the columns of $\mathbf{Z}$ are independent of those of $\mathbf{X}$.

**(*ii*) *Solving the Normal Equations.*** The normal equations for $\mathbf{a}^\circ$ and $\mathbf{b}^\circ$ are, from (6),

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{a}^\circ \\ \mathbf{b}^\circ \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}. \tag{7}$$

Suppose $(\mathbf{X}'\mathbf{X})^-$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Then the first equation of (7) gives

$$\begin{aligned}\mathbf{a}^\circ &= (\mathbf{X}'\mathbf{X})^-(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\mathbf{b}^\circ) \\ &= (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{b}^\circ \\ &= \mathbf{a}^* - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{b}^\circ\end{aligned} \tag{8}$$

where

$$\mathbf{a}^* = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$$

is the solution to the normal equation without the covariate. Substituting for $\mathbf{a}^\circ$ into (7) gives the solution for $\mathbf{b}^\circ$,

$$\mathbf{b}^\circ = \{\mathbf{Z}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{Z}\}^-\mathbf{Z}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{y}. \tag{9}$$

Again, the superscripted minus sign designates a generalized inverse. Substitution of (9) into (8) gives $\mathbf{a}^\circ$ explicitly. Solutions (8) and (9) are exactly the same results, as would be obtained by using the expression for a generalized inverse given in Section 7 of Chapter 1 (see Exercise 13).

We should note several features of (9). First, although $(\mathbf{X}'\mathbf{X})^-$ is not unique, it enters into $\mathbf{b}^\circ$ only in the form $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$. This is invariant to whatever generalized inverse of $\mathbf{X}'\mathbf{X}$ is used for $(\mathbf{X}'\mathbf{X})^-$. Thus the non-full-rank property of $\mathbf{X}$ does not of itself, lead to more than one solution for $\mathbf{b}^\circ$. Suppose we use $\mathbf{P}$ for

$$\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' \tag{10}$$

By Theorem 10 of Chapter 1, $\mathbf{P}$ is symmetric and idempotent. Then (9) can be written as $\mathbf{b}^\circ = (\mathbf{Z}'\mathbf{P}\mathbf{Z})^-\mathbf{Z}'\mathbf{P}\mathbf{y}$. Symmetry and idempotency of $\mathbf{P}$ ensure that $\mathbf{Z}'\mathbf{P}\mathbf{Z}$ and $\mathbf{P}\mathbf{Z}$ have the same rank. Furthermore, the properties of $\mathbf{X}$ and $\mathbf{Z}$ given below (6) guarantee that $\mathbf{P}\mathbf{Z}$ has full-column rank and hence $\mathbf{Z}'\mathbf{P}\mathbf{Z}$ is non-singular (see Exercise 13). Therefore, $\mathbf{b}^\circ$ is the sole solution.

$$\mathbf{b}^\circ = \hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{P}\mathbf{Z})^-\mathbf{Z}'\mathbf{P}\mathbf{y}. \tag{11}$$

**(*iii*) *Estimability.*** Consideration of the expected value of $\hat{\mathbf{b}}$ of (11) and of $\mathbf{a}^\circ$ of (8) show that $\mathbf{b}$ is estimable and that $\lambda'\mathbf{a}$ is estimable when $\lambda' = \mathbf{t}'\mathbf{X}$ for some $\mathbf{t}'$. That means that $\mathbf{b}$ is always estimable and $\lambda'\mathbf{a}$ is estimable whenever it is estimable for the model that has no covariates. (See Exercise 13.)

**(*iv*) *A Model for Handling the Covariates.*** The estimator $\hat{\mathbf{b}}$ shown in (11) is the b.l.u.e. of $\mathbf{b}$ in the model (6). By the form of (11), it is also the b.l.u.e. of $\mathbf{b}$ in the model having the equation

$$\mathbf{y} = \mathbf{P}\mathbf{Z}\mathbf{b} + \mathbf{e}. \tag{12}$$

This, we shall see provides a convenient method for estimating $\mathbf{b}$.

Recall that in fitting a model of the form $\mathbf{y} = \mathbf{Xa} + \mathbf{e}$, the vector of estimated expected values $\hat{\mathbf{y}}$ corresponding to the vector of observed values $\mathbf{y}$ is $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (equation (10), Chapter 5). Therefore, the vector of residuals, that is, the vector of deviations of the observed values from their corresponding estimated values is

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}.$$

This becomes, using (10),

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{P}\mathbf{y}.$$

Thus, $\mathbf{Py}$ is the vector of $y$-residuals after fitting the model $\mathbf{y} = \mathbf{Xa} + \mathbf{e}$. Similarly, if the $j$th column of $\mathbf{Z}$ is $\mathbf{z}_j$, the $j$th column of $\mathbf{PZ}$ in (12) is $\mathbf{Pz}_j$, the vector of $\mathbf{z}_j$ residuals after fitting the model[1] $\mathbf{z}_j = \mathbf{Xa} + \mathbf{e}$. Thus with

$$\mathbf{Z} = \{\mathbf{z}_j\} \quad \text{for} \quad j = 1, 2, \ldots, q,$$

we write $\mathbf{R}_z$ for $\mathbf{PZ}$ and have $\mathbf{R}_z$ as the matrix of residuals. Thus,

$$\mathbf{R}_z = \mathbf{PZ} = \{\mathbf{P}\mathbf{z}_j\} = \{\mathbf{z}_j - \hat{\mathbf{z}}_j\} = \{\mathbf{z}_j - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{z}_j\}. \tag{13}$$

Hence, the model (12) is equivalent to the model

$$\mathbf{y} = \mathbf{R}_\mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{14}$$

and $\hat{\mathbf{b}}$ of (11) is

$$\hat{\mathbf{b}} = (\mathbf{R}_z'\mathbf{R}_z)^{-1}\mathbf{R}_z'\mathbf{y}.$$

The matrix $\mathbf{R}_z$ has the same order as $\mathbf{Z}$. Its columns are columns of residuals given in (13). The matrix of sums of squares and products of $z$-residuals is $\mathbf{R}_\mathbf{z}'\mathbf{R}_\mathbf{z}$. The vector of sums and products of $z$-residuals and the $y$-observations is $\mathbf{R}_\mathbf{z}'\mathbf{y}$.

**(*v*) *Analyses of Variance.*** The reduction in sum of squares for fitting a linear model is the inner product of a solution vector and the vector of the right-hand sides of the normal equations (e.g., equation (14) of Chapter 5). Thus, from (7), (8), and (11), the reduction in the sum of squares for fitting the model is

$$R(\mathbf{a}, \mathbf{b}) = \mathbf{a}^{\circ\prime}\mathbf{X}'\mathbf{y} + \hat{\mathbf{b}}\mathbf{Z}'\mathbf{y}.$$

[1] S. R. Searle is grateful for discussions with N. S. Urquhart.

In the notation $R(\mathbf{a}, \mathbf{b})$, $\mathbf{b}$ emphasizes the fitting of a vector of coefficients pertaining to the covariates and $\mathbf{a}$ represents the factor and interactions part of the model, including $\boldsymbol{\mu}$. Upon substitution for $\mathbf{a}^\circ$ and $\hat{\mathbf{b}}$ from (8) and (11) making use of (10), $R(\mathbf{a}, \mathbf{b})$ reduces to

$$\begin{aligned} R(\mathbf{a}, \mathbf{b}) &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + \mathbf{y}'\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}\mathbf{y} \\ &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + \mathbf{y}'\mathbf{R}_{\mathrm{Z}}(\mathbf{R}_{\mathrm{z}}'\mathbf{R}_{\mathrm{Z}})^{-1}\mathbf{R}_{\mathrm{Z}}'\mathbf{y} \end{aligned}$$

This is the sum of two reductions. The first one is

$$R(\mathbf{a}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}, \qquad \text{due to fitting } \mathbf{y} = \mathrm{X}\mathbf{a} + \mathbf{e}.$$

The second one is

$$\mathrm{SSRB} = \mathbf{y}'\mathbf{R}_{\mathrm{z}}(\mathbf{R}_{\mathrm{z}}'\mathbf{R}_{\mathrm{z}})^{-1}\mathbf{R}_{\mathrm{z}}'\mathbf{y} = \hat{\mathbf{b}}\mathbf{R}_{\mathrm{z}}'\mathbf{y}, \qquad \text{due to fitting } \mathbf{y} = \mathbf{R}_{\mathrm{z}}\mathbf{b} + \mathbf{e}.$$

Putting these two expressions together, we have

$$R(\mathbf{a}, \mathbf{b}) = R(\mathbf{a}) + \mathrm{SSRB}.$$

Consequently,

$$R(\mathbf{b}|\mathbf{a}) = R(\mathbf{a}, \mathbf{b}) - R(\mathbf{a}) = \mathrm{SSRB} = \hat{\mathbf{b}}'\mathbf{R}_{\mathrm{z}}'\mathbf{y}.$$

Thus, SSRB is the reduction in the sum of squares attributable to fitting the covariates, having already fitted the factor and interactions part of the model.

Distributional properties of $R(\mathbf{a})$ and $R(\mathbf{b}|\mathbf{a})$, based on the usual normality assumptions, come from Theorems 5 and 6 of Chapter 2. The idempotency of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ and of $\mathbf{R}_{\mathrm{z}}(\mathbf{R}_{\mathrm{z}}'\mathbf{R}_{\mathrm{z}})^{-1}\mathbf{R}_{\mathrm{z}}'$ give

$$\frac{R(\mathbf{a})}{\sigma^2} \sim \chi^{2'}[r(\mathbf{X}), \lambda_a]$$

with

$$\lambda_a = \frac{\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} + 2\mathbf{a}'\mathbf{X}'\mathbf{Z}\mathbf{b} + \mathbf{b}'\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{b}}{2\sigma^2}$$

and

$$\frac{R(\mathbf{b}|\mathbf{a})}{\sigma^2} \sim \chi^{2'}\left[r(\mathbf{Z}), \frac{\mathbf{b}'\mathbf{R}_{\mathrm{z}}'\mathbf{R}_{\mathrm{z}}\mathbf{b}}{2\sigma^2}\right].$$

**TABLE 8.3a Analysis of Variance for Fitting Covariates (b) After Factors and Interactions (a) in the Covariance Model $\mathbf{y} = \mathbf{Xa} + \mathbf{Zb} + \mathbf{e}$**

| Source of Variation | d.f. | Sum of Squares[a] |
|---|---|---|
| Factors and Interaction | $r(\mathbf{X})$ | $R(\mathbf{a}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ |
| Mean | 1 | $R(\mu) = N\bar{y}^2$ |
| Factors and interactions (after the mean) | $r(\mathbf{X}) - 1$ | $R(\mathbf{a}\|\mu) = R(\mathbf{a}) - R(\mu)$ |
| Covariates (after factors and interactions) | $r(\mathbf{Z})$ | $R(\mathbf{b}\|\mathbf{a}) = \text{SSRB} = \mathbf{y}'\mathbf{R}_Z(\mathbf{R}_Z'\mathbf{R}_Z)^{-1}\mathbf{R}_Z'\mathbf{y}$ |
| Residual error | $N - r(\mathbf{X}) - r(\mathbf{Z})$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - R(\mathbf{a}) - \text{SSRB}$ |
| Total | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ |

[a] $\mathbf{R}_Z$ is the matrix of residuals in (13).

We show that $R(\mathbf{a})$ and $R(\mathbf{b}|\mathbf{a})$ are distributed independently. Recall that $\mathbf{R}_Z = \mathbf{PZ}$. Furthermore, by the definition of $\mathbf{P}$ in (10), $\mathbf{X}'\mathbf{P} = \mathbf{0}$. It follows that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{R}_Z(\mathbf{R}_z'\mathbf{R}_Z)^{-1}\mathbf{R}_z' = \mathbf{0}$.

Hence, $R(\mathbf{a})$ and $R(\mathbf{b}|\mathbf{a})$ are independent random variables. The reader can show in Exercise 11 that $R(\mathbf{a})$ and $R(\mathbf{b}|\mathbf{a})$ is also independent of

$$\text{SSE} = \mathbf{y}'\mathbf{y} - R(\mathbf{a}, \mathbf{b}) = \mathbf{y}'\mathbf{y} - R(\mathbf{a}) - \text{SSRB}.$$

The statistic SSE also has a $\chi^2$-distribution $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{N-r(X)-r(z)}$.

These sums of squares are summarized in Table 8.3a. Mean squares and $F$-statistics follow in the usual way.

The unbiased estimator of $\sigma^2$ which we can derive from Table 8.3a is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N - r(\mathbf{X}) - r(\mathbf{Z})}.$$

An alternative to the analysis of variance shown in Table 8.3a is to fit the covariates before the factors and interactions instead of after them, as is done there. This necessitates calculating $R(\mathbf{b}|\mu) = R(\mu, \mathbf{b}) - R(\mu)$. to do this, we need $R(\mu, \mathbf{b})$, the reduction in the sum of squares due to fitting the model

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Zb} + \mathbf{e}.$$

Of course, this is simply an intercept regression model. The estimators of the parameters $\mu$ and $\mathbf{b}$ are

$$\tilde{\mathbf{b}} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y} \quad \text{and} \quad \hat{\mu} = \bar{y} - \tilde{\mathbf{b}}\bar{\mathbf{z}}$$

as in (41) and (42) of Chapter 3. In $\tilde{\mathbf{b}}$, $\mathcal{X}'\mathcal{X}$ is the matrix of corrected sums and squares and products of the observed $z$'s. Furthermore, $\mathcal{X}'\mathbf{y}$ is the vector of corrected sums

of products of the $z$'s and the $y$'s. Then, $R(\mathbf{b}|\mu)$ that we need here is $\text{SSR}_m$ of (87) in Section 4f of Chapter 3. As a result,

$$R(\mathbf{b}|\mu) = \mathbf{y}'\mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y}.$$

This reduction sum of squares is for fitting covariates after the mean.

In addition, we need that for fitting the factors and interactions after the mean and covariates:

$$R(\mathbf{a}|\mu, \mathbf{b}) = R(\mathbf{a}, \mathbf{b}) - R(\mu, \mathbf{b}),$$

remembering that $\mathbf{a}$ in this notation includes $\mu$. On using $R(\mathbf{a}) + \text{SSRB}$ for $R(\mathbf{a}, \mathbf{b})$ as derived in establishing Table 8.3a and $R(\mathbf{b}|\mu) + R(\mu) = R(\mu, \mathbf{b})$, we have

$$\begin{aligned} R(\mathbf{a}|\mu, \mathbf{b}) &= R(\mathbf{a}) + \text{SSRB} - R(\mathbf{b}|\mu) - R(\mu) \\ &= R(\mathbf{a}|\mu) + \text{SSRB} - R(\mathbf{b}|\mu). \end{aligned}$$

These calculations are summarized in Table 8.3.

In both Tables 8.3a and 8.3b, the terms $R(\mu)$ and $R(\mathbf{a}|\mu)$ are those familiarly calculated in the no-covariate model $\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}$.

***(vi) Tests of Hypotheses.*** The distributional properties of $R(\mathbf{b}|\mathbf{a})$ and SSE indicate, from (14), that in Table 8.3a,

$$F(\mathbf{b}|\mathbf{a}) = \frac{R(\mathbf{b}|\mathbf{a})/r(\mathbf{Z})}{\text{SSE}/[N - r(\mathbf{X}) - r(\mathbf{Z})]}$$

tests the hypothesis $H$: $\mathbf{b} = \mathbf{0}$.

The hypothesis $H$: $\mathbf{K}'\mathbf{a} = \mathbf{m}$ is testable provided that $\mathbf{K}'\mathbf{a}$ is testable. If this is the case, we can test the hypothesis in the usual manner given by equation (71) of Chapter 5. To use that equation with the solutions $\mathbf{a}^\circ$ and $\mathbf{b}^\circ$ given in (8) and (9), we need the generalized inverse of the partitioned matrix shown in (7). From

**TABLE 8.3b Analysis of Variance for Fitting Factors and Interactions (a) After Covariates (b) in the Covariance Model $\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{Z}\mathbf{b} + \mathbf{e}$**

| Source of Variation | d.f. | Sum of Squares[a] |
|---|---|---|
| Mean | 1 | $R(\mu) = N\bar{y}^2$ |
| Covariates (after mean) | $r(\mathbf{Z})$ | $R(\mathbf{b}\|\mu) = \mathbf{y}'\mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y}$ |
| Factors and Interactions (after mean and covariates) | $r(\mathbf{X}) - 1$ | $R(\mathbf{a}\|\mu, \mathbf{b}) = R(\mathbf{a}\|\mu) + \text{SSRB} - R(\mathbf{b}\|\mu)$ |
| Residual error | $N - r(\mathbf{X}) - r(\mathbf{Z})$ | $\text{SSE} = \mathbf{y}'\mathbf{y} - R(\mathbf{a}) - \text{SSRB}$ |
| Total | $N$ | $\text{SST} = \mathbf{y}'\mathbf{y}$ |

[a] $R(\mathbf{a}|\mu)$ and SSRB are given in Table 8.3a.

equation (56) of Chapter 1, this generalized inverse is (see Exercise 33 in Chapter 1).

$$\mathbf{G} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^{-} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z} \\ \mathbf{I} \end{bmatrix} (\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1} [-(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z} \quad \mathbf{I}]. \tag{15}$$

Writing the hypothesis $H$: $\mathbf{K}'\mathbf{a} = \mathbf{m}$ as

$$H\colon [\mathbf{K}' \quad \mathbf{0}] \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \mathbf{m}$$

it will be found that the numerator of $F(H)$ reduces to

$$Q = (\mathbf{K}'\mathbf{a}^{\circ} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K} + \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-1} \times (\mathbf{K}'\mathbf{a}^{\circ} - \mathbf{m}).$$

We now show that testing $H$: $\mathbf{K}'\mathbf{a} = \mathbf{0}$ in the no-covariance model has the same numerator sum of squares as does testing $H$: $\mathbf{K}'[\mathbf{a} + (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{b}] = \mathbf{0}$ in the covariance model. The solution vector for $\mathbf{a}$ in the no-covariance model is $\mathbf{a}^{*} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$. From $Q$ of Table 5.9, the numerator sum of squares for testing $H$: $\mathbf{K}'\mathbf{a} = 0$ in the no-covariance model is therefore,

$$Q = \mathbf{a}^{*'}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-}\mathbf{K}'\mathbf{a}^{*}. \tag{16}$$

In the covariance model, consider the hypothesis

$$H\colon \mathbf{K}'[\mathbf{a} + (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{b}] = \mathbf{0}. \tag{17}$$

This can be written as

$$H\colon \mathbf{K}'[\mathbf{I} \quad (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}] \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \mathbf{0} \quad \text{or as} \quad \mathbf{M}' \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \mathbf{0}$$

with

$$\mathbf{M}' = \mathbf{K}'[\mathbf{I} \quad (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}] \tag{18}$$

We may test this hypothesis using an $F$-statistic having numerator sum of squares (see Table 5.9)

$$Q_c = [\mathbf{a}^{\circ'} \quad \hat{\mathbf{b}}'']\mathbf{M}(\mathbf{M}'\mathbf{G}\mathbf{M})^{-1}\mathbf{M}' \begin{bmatrix} \mathbf{a}^{\circ} \\ \hat{\mathbf{b}} \end{bmatrix}.$$

However, from (15) and (18), $\mathbf{M}'\mathbf{G}\mathbf{M} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}$, and from (8) $[\mathbf{a}^{\circ\prime} \quad \hat{\mathbf{b}}']\mathbf{M} = \mathbf{a}^{*\prime}\mathbf{K}$.

Thus, $Q_c$ becomes

$$\begin{aligned} Q_c &= (\mathbf{K}'\mathbf{a}^*)'(\mathbf{K}'(\mathbf{X}'\mathbf{X})\mathbf{K}]^{-1}\mathbf{K}'\mathbf{a}^* \\ &= Q \text{ of (16).} \end{aligned}$$

Hence, the numerator sum of squares for testing $H$: $\mathbf{K}'\mathbf{a} = \mathbf{0}$ in the no-covariance model is also the numerator sum of squares for testing

$$H\text{: } \mathbf{K}'[\mathbf{a} + (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{b}] = \mathbf{0}$$

in the covariance model. This hypothesis appears to be dependent on $(\mathbf{X}'\mathbf{X})^-$. This is not the case, because $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ for some $\mathbf{T}$, since we assume that $H$: $\mathbf{K}'\mathbf{a} = \mathbf{0}$ is testable.

**(*vii*) *Summary.*** We can summarize the preceding development of the analysis of covariance model

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

as follows. First fit

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}.$$

Calculate

$$\mathbf{a}^* = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} \quad \text{and} \quad R(\mathbf{a}) = \mathbf{a}^{*'}\mathbf{X}'\mathbf{y}. \tag{19}$$

Then for each column of $\mathbf{Z}$, $\mathbf{z}_j$ say, fit

$$\mathbf{z}_j = \mathbf{X}\mathbf{a} + \mathbf{e}.$$

Calculate the $\mathbf{z}_j$-residual vector

$$\mathbf{z}_j - \hat{\mathbf{z}}_j = \mathbf{z}_j - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{z}_j$$

and the matrix of these residuals

$$\mathbf{R}_z = \{\mathbf{z}_j - \hat{\mathbf{z}}_j\} \quad \text{for} \quad j = 1, 2, \ldots, q. \tag{20}$$

Fit

$$\mathbf{y} = \mathbf{R}_z\mathbf{b} + \mathbf{e} \tag{21}$$

and calculate

$$\hat{\mathbf{b}} = (\mathbf{R}_z'\mathbf{R}_z)^{-1}\mathbf{R}_z'\mathbf{y} \tag{22}$$

and

$$R(\mathbf{b}|\mathbf{a}) = \hat{\mathbf{b}}'\mathbf{R}_z'\mathbf{y}.$$

The solution vector for the covariance matrix is then

$$\begin{bmatrix} \mathbf{a}^\circ \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{a}^* - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\hat{\mathbf{b}} \\ \hat{\mathbf{b}} \end{bmatrix}. \tag{23}$$

From (15), the dispersion matrices of these solutions are

$$\begin{aligned} \text{var}(\mathbf{a}^\circ) &= [(\mathbf{X}'\mathbf{X})^- + (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}(\mathbf{R}_z'\mathbf{R}_z)^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-]\sigma^2, \\ \text{var}(\hat{\mathbf{b}}) &= (\mathbf{R}_z'\mathbf{R}_z)^{-1}\sigma^2 \end{aligned} \tag{24}$$

and

$$\text{cov}(\mathbf{a}^\circ, \hat{\mathbf{b}}) = -(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}(\mathbf{R}_z'\mathbf{R}_z)^{-1}\sigma^2.$$

In contrast to fitting an ordinary factors-and-interaction model, the clue to the calculations for a covariance model is the derivation of $\mathbf{R}_z$. Furthermore, the calculation of each column of $\mathbf{R}_z$, from the corresponding column of $\mathbf{Z}$ depends solely on the particular factor-and-interactions model being used. No matter what the nature of the covariates, $\mathbf{X}$ is the same for any specific factors-and-interactions model. The matrix $\mathbf{X}$ is what determines the derivation of $\mathbf{R}_z$ from $\mathbf{Z}$. When considering the same covariates in different ways for the same factors-and-interactions model, the corresponding $\mathbf{Z}$ matrices will be different, but the mode of calculating $\mathbf{R}_z$ is always the same. The columns of $\mathbf{R}_z$ are always the vectors of residuals obtained after fitting the no-covariates model to each column of $\mathbf{Z}$. This is illustrated in the examples that follow.

### b. The One-Way Classification

***(i) A Single Regression.*** An adaption of equation (23) in Chapter 6 gives the equation for a covariance model in the one-way classification as

$$y_{ij} = \mu + \alpha_i + bz_{ij} + e_{ij} \tag{25}$$

for $i = 1, 2, \ldots, c$ and $j = 1, 2, \ldots, n_i$. In this model, $\mu$ and the $\alpha_i$'s are the elements of $\mathbf{a}$ of (6). The scalar $b$ is the sole element of $\mathbf{b}$ of (6). The matrix $\mathbf{Z}$ of (6) is a column

vector of observed values $z_{ij}$ of the covariate, with

$$\mathbf{z}' = [z_{11} \quad z_{12} \quad \cdots \quad z_{1n_1} \quad \cdots \quad z_{i1} \quad z_{i2} \quad \cdots \quad z_{in_i} \quad \cdots \quad z_{c1} \quad z_{c2} \quad \cdots \quad z_{cn_c}], \tag{26}$$

corresponding to the vector of y observations defined in (26) of Chapter 6.

Fitting the no-covariate form of (25) amounts to fitting the one-way classification model $y_{ij} = \mu + \alpha_i + e_{ij}$ discussed in Section 2 of Chapter 6. There in equation (31), we see that a solution vector for $\mathbf{a}^*$ of (19) is

$$\mathbf{a}^* = \begin{bmatrix} \mu^* \\ \{\alpha_i^*\} \end{bmatrix} = \begin{bmatrix} 0 \\ \{\bar{y}_{i.}\} \end{bmatrix} \quad \text{for} \quad i = 1, \ldots, c. \tag{27}$$

From (37) of Section 2d of Chapter 6, it follows that

$$R(\mathbf{a}) = \sum_{i=1}^{c} \frac{y_{i.}^2}{n_i}. \tag{28}$$

Furthermore, the residual corresponding to $y_{ij}$ is

$$y_{ij} - \hat{y}_{ij} = y_{ij} - \mu^* - \alpha_i^* = y_{ij} - \bar{y}_{i.}.$$

Then the vector of residuals is

$$\mathbf{y} - \hat{\mathbf{y}} = \{y_i - \bar{y}_{i.}\mathbf{1}_{n_i}\} = \left\{ \begin{bmatrix} y_{i1} - \bar{y}_{i.} \\ y_{i2} - \bar{y}_{i.} \\ \vdots \\ y_{in_i} - \bar{y}_{i.} \end{bmatrix} \right\} \quad \text{for } i = 1, \ldots, c. \tag{29}$$

In fitting (25), $\mathbf{Z}$ of the general model (6) is $\mathbf{z}$ of (26). Then $\mathbf{R}_z$ of (20) is a vector.

Analogous to (29), we have

$$\mathbf{R}_z = \mathbf{z} - \hat{\mathbf{z}} = \{\mathbf{z}_i - \bar{z}_{i.}\mathbf{1}_{n_i}\} \quad \text{for} \quad i = 1, 2, \ldots, c.$$

Therefore, for $\hat{\mathbf{b}}$ of (22)

$$\mathbf{R}_z'\mathbf{R}_z = \sum_{i=1}^{c}\sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2 = \sum_{i=1}^{c}\left(\sum_{j=1}^{n_i} z_{ij}^2 - n_i\bar{z}_{i.}^2\right) \tag{30a}$$

and

$$\mathbf{R}_z'\mathbf{y} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})y_{ij} = \sum_{i=1}^{c}\left(\sum_{j=1}^{n_i} y_{ij}z_{ij} - n_i\bar{y}_{i.}\bar{z}_{i.}\right). \tag{30b}$$

Thus,

$$\hat{b} = \frac{\sum_{i=1}^{c}\left(\sum_{j=1}^{n_i} y_{ij}z_{ij} - n_i\bar{y}_{i\cdot}\bar{z}_{i\cdot}\right)}{\sum_{i=1}^{c}\left(\sum_{j=1}^{n_i} z_{ij}^2 - n_i\bar{z}_{i\cdot}^2\right)}. \tag{31}$$

With the value of $\hat{b}$ in (31), we can calculate $\mathbf{a}^\circ$ from (23) as

$$\mathbf{a}^\circ = \mathbf{a}^* - \hat{b}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{z};$$

that is,

$$\begin{bmatrix} \mu^\circ \\ \{\alpha_i^\circ\} \end{bmatrix} = \begin{bmatrix} 0 \\ \{\bar{y}_{i\cdot}\} \end{bmatrix} - \hat{b}\begin{bmatrix} 0 \\ \{\bar{z}_{i\cdot}\} \end{bmatrix} = \begin{bmatrix} 0 \\ \{\bar{y}_{i\cdot} - \hat{b}\bar{z}_{i\cdot}\} \end{bmatrix} \quad \text{for} \quad i = 1, \ldots, c. \tag{32}$$

The solution $a_i^\circ = \bar{y}_i - \hat{b}\bar{z}_i$ is often referred to as an *adjusted mean.* It is the class mean $\bar{y}_{i\cdot}$ adjusted by the class mean of the covariate, using the estimate $\hat{b}$ to make the adjustment.

Examination of (31) and (32) reveals the relationship of these results to ordinary regression analysis. In (31), the numerator of $\hat{b}$ is a sum of terms, each of which is the numerator for estimating the within class regression of $y$ on $z$. Likewise, the denominator of $\hat{b}$ is the sum of the denominators of these within class regression estimators. Thus, $\hat{b}$ is usually referred to as the *pooled within-class regression estimator.* Moreover, each element of (32)—other than the initial zero—is the within class intercept estimator using $\hat{b}$ of (31).

The basic calculations for the analysis of variance for fitting the model $E(\mathbf{y}) = \mathbf{Xa}$ in the case of a one-way classification are, as in Section 2d of Chapter 6

$$\mathrm{SSR}_{yy} = \sum_{i=1}^{c} n_i\bar{y}_{\cdot}^2, \quad \mathrm{SSE}_{yy} = \mathrm{SST}_{yy} - \mathrm{SSR}_{yy} \quad \text{and} \quad \mathrm{SST}_{yy} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2.$$

We can also calculate

$$\mathrm{SSM}_{yy} = N\bar{y}^2, \quad \mathrm{SSR}_{m,yy} = \mathrm{SSR}_{yy} - \mathrm{SSM}_{yy}$$

and

$$\mathrm{SST}_{m,yy} = \mathrm{SST}_{yy} - \mathrm{SSM}_{yy}.$$

The subscript $yy$ in these expressions emphasizes that they are functions of squares of the $y$-observations. We can also calculate similar functions of the $z$-observations, and of cross products of the $y$'s and $z$'s. The basic calculations include

$$\mathrm{SSR}_{yz} = \sum_{i=1}^{c} n_i \bar{y}_{i.} \bar{z}_{i.}, \ \mathrm{SSE}_{yz} = \mathrm{SST}_{yz} - \mathrm{SSR}_{yz}, \quad \text{and} \quad \mathrm{SST}_{yz} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} y_{ij}^2.$$

We also have that

$$\mathrm{SSM}_{yz} = N\bar{y}\bar{z}, \quad \mathrm{SSR}_{m,yz} = \mathrm{SSR}_{yz} - \mathrm{SSM}_{yz},$$

and

$$\mathrm{SST}_{m,yz} = \mathrm{SST}_{yz} - \mathrm{SSM}_{yz}.$$

(We do not show explicit expressions for the $z$'s because they are exactly of the same form as the $y$'s.) We find these expressions useful in what follows.

First, $R(\mathbf{a})$, which for (25) is the reduction due to fitting $\mu$ and the $\alpha$'s is from (28)

$$R(\mu, \alpha) = R(\mathbf{a}) = \mathrm{SSR}_{yy}.$$

Second, from (31),

$$\hat{b} = \frac{\mathrm{SSE}_{yz}}{\mathrm{SSE}_{zz}}. \tag{33}$$

From (22), (30), and (33), we have

$$R(b|\mu, \alpha) = R(\mathbf{b}|\mathbf{a}) = \frac{(\mathrm{SSE}_{yz})^2}{\mathrm{SSE}_{zz}}. \tag{34}$$

Hence, the analysis of variance of Table 8.3 can be rewritten in the form of Table 8.4.

In Table 8.4, the estimated residual variance is

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{N - c - 1}.$$

We test the hypothesis that the regression slope is zero, that is, $H$: $b = 0$ using the $F$-statistic,

$$F(b) = \frac{R(b|\mu, \alpha)}{\hat{\sigma}^2}, \tag{35}$$

**TABLE 8.4a Analysis of Variance for Fitting the Covariate after the Class Effects in the One-Way Classification Covariance Model $y_{ij} = \mu + \alpha_i + bz_{ij} + e_{ij}$**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = \text{SSM}_{yy}$ |
| $\alpha$-classes (after mean) | $c - 1$ | $R(\alpha\|\mu) = \text{SSR}_{m,yy}$ |
| Covariate (pooled within-class regression) | 1 | $R(b\|\mu, \alpha) = \dfrac{(\text{SSE}_{yz})^2}{\text{SSE}_{zz}}$ |
| Residual error | $N - c - 1$ | $\text{SSE} = \text{SSE}_{yy} - R(b\|\mu, \alpha)$ |
| Total | $N$ | $\text{SST}_{yy}$ |

with 1 and $N-c-1$ degrees of freedom. For the no-covariate model the $F$-statistic with $R(\alpha|\mu)$ in its numerator tests the hypothesis $H$: all $\alpha$'s equal (See Section 2f(iii) of Chapter 6). From (17), the corresponding $F$-statistic in Table 8.4a tests the hypothesis

$$H\text{: } \alpha_i + b\bar{z}_{i.} \text{ equal for all } i. \tag{36}$$

The $b\bar{z}_{i.}$ in (36) are derived from $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}\mathbf{b}$ of (17) in the same that $\mathbf{a}^{\circ}$ of (32) was derived. This hypothesis represents equality of the $\alpha$'s adjusted for the observed $z$'s.

To derive the equivalent of Table 8.3b for the one-way classification covariance model, notice first that whenever there is only a single vector as $\mathbf{Z}$, then in Table 8.3b

$$\mathbf{y}'\mathcal{X} = \text{SST}_{m,yz} \quad \text{and} \quad \mathcal{X}'\mathcal{X} = \text{SST}_{m,zz}.$$

Hence,

$$R(b|\mu) = \frac{\left(\text{SST}_{m,yz}\right)^2}{\text{SST}_{m,zz}}.$$

As a result, Table 8.3b simplifies to Table 8.4b.

**TABLE 8.4b Analysis of Variance for Fitting the Class Effects After the Covariate in the One-Way Classification Covariance Model $y_{ij} = \mu + \alpha_i + bz_{ij} + e_{ij}$**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = \text{SSM}_{yy}$ |
| Covariate (after mean) | 1 | $R(b\|\mu) = \dfrac{\left(\text{SST}_{m,yz}\right)^2}{\text{SST}_{m,zz}}$ |
| $\alpha$-classes (after mean and covariates) | $c - 1$ | $R(\alpha\|\mu, b) = \text{SSR}_{m,yy} + \dfrac{\left(\text{SSE}_{y,z}\right)^2}{\text{SSE}_{zz}} - \dfrac{\left(\text{SST}_{m,yz}\right)^2}{\text{SST}_{m,zz}}$ |
| Residual error | $N - c - 1$ | $\text{SSE} = \text{SSE}_{yy} - R(b\|\mu, \alpha)$ |
| Total | $N$ | $\text{SST}_{yy}$ |

We can derive $F$-statistic for testing

$$H\text{: } \alpha_i \text{ equal for all } i$$

by writing the hypothesis as $\mathbf{K}'\mathbf{a} = \mathbf{0}$ and using the general result for $Q$ given below (15). A possible value for $\mathbf{K}'$ would be $\mathbf{K}' = [0\mathbf{1}_{c-1} \quad \mathbf{1}_{c-1} \quad -\mathbf{I}_{c-1}]$.

An easier development would be to consider the reduced model arising from the hypothesis itself, namely

$$y_{ij} = (\mu + \alpha) + bz_{ij} + e_{ij}. \tag{37}$$

This is a model for simple regression, for which the estimator of $b$ is, from equation (14) of Chapter 3

$$\tilde{b} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}z_{ij} - N\bar{y}\bar{z}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} z_{ij}^2 - N\bar{z}^2} = \frac{\mathrm{SST}_{m,yz}}{\mathrm{SST}_{m,zz}}. \tag{38}$$

The reduction in the sum of squares for fitting (37) is therefore, using Table 3.3 of Chapter 3

$$\begin{aligned} R(\mu, b) &= N\bar{y}^2 + \tilde{b}\mathrm{SST}_{m,yz} \\ &= \mathrm{SSM}_{yy} + \frac{\left(\mathrm{SST}_{m,yz}\right)^2}{\mathrm{SST}_{m,zz}}. \end{aligned} \tag{39}$$

The full model is (25), with the reduction in sum of squares being, from Table 8.4a

$$R(\mu, \alpha, b) = \mathrm{SSM}_{yy} + \mathrm{SSR}_{m,yy} + R(b|\mu, \alpha). \tag{40}$$

The $F$-statistic for testing $H$: all $\alpha$'s equal in the model (25) has numerator

$$Q = R(\mu, \alpha, b) - R(\mu, b). \tag{41}$$

Using (34), (38), and (39), this becomes $Q = R(\alpha|\mu, b)$ of Table 8.4b. Tables similar to 8.4a and 8.4b are to be found in many places; for example, Federer (1955, p. 486), Graybill (1961, pp. 385 and 393), and Montgomery ((2005) pp. 577–578).

**(*ii*) *Example.***

**Example 1 Relationship between Number of Children and Investment Index for Men of Different Levels of Education** In Section 2 of Chapter 6, we considered an example that compared the investment indices for men of three different levels of

**TABLE 8.5 Investment Index and Number of Children for Seven Men**

| High School Incomplete | | High School Graduate | | College Graduate | |
|---|---|---|---|---|---|
| Index, $y_{1j}$ | Children, $z_{1j}$ | Index, $y_{2j}$ | Children, $z_{2j}$ | Index, $y_{3j}$ | Children, $z_{3j}$ |
| 74 | 3 | 76 | 2 | 85 | 4 |
| 68 | 4 | 80 | 4 | 93 | 6 |
| 77 | 2 | | | | |
| 219 | 9 | 156 | 6 | 178 | 10 |

education. The levels were high school incomplete, high school graduate and college graduate. We consider this example again introducing a covariate, the number of children in each family. We consider hypothetical data in Table 8.5 where the $y$-values (investment index) are the same as in Table 6.1.

We may calculate the following basic sums of squares from Table 8.5.

| | | |
|---|---|---|
| $\text{SSR}_{yy} = 43{,}997$ | $\text{SSR}_{zz} = 95$ | $\text{SSR}_{yz} = 2015$ |
| $\text{SSE}_{yy} = 82$ | $\text{SSE}_{zz} = 6$ | $\text{SSE}_{yz} = 3$ |
| $\text{SST}_{yy} = 44{,}079$ | $\text{SST}_{zz} = 101$ | $\text{SST}_{yz} = 2018$ |
| $\text{SSM}_{yy} = 43{,}687$ | $\text{SSM}_{zz} = 89.2857$ | $\text{SSM}_{yz} = 1975$ |

We shall use these numbers in the ensuing calculations.

From (33), we obtain the pooled regression estimate

$$\hat{b} = \frac{3}{6} = \frac{1}{2} = 0.5 \tag{42}$$

For $\mathbf{a}^\circ$ of (32), we need $\mathbf{a}^*$ of (27). From (34) of Chapter 6, we have

$$\mathbf{a}^{*\prime} = [0 \quad 73 \quad 78 \quad 89]. \tag{43}$$

Hence from (32) and Table 8.5,

$$\mathbf{a}^\circ = \begin{bmatrix} 0 \\ 73 \\ 78 \\ 89 \end{bmatrix} - 0.5 \begin{bmatrix} 0 \\ 3 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 71.5 \\ 76.5 \\ 86.5 \end{bmatrix}. \tag{44}$$

The analysis of variance in Table 8.4a uses

$$\begin{aligned} R(\mu) &= \text{SSM}_{yy} = 43{,}687 \\ R(\mu, \alpha) &= \text{SSR}_{yz} = 43{,}997 \\ R(b|\mu, \alpha) &= \text{SSRB} = \frac{3^2}{6} = 1.5 \end{aligned} \tag{45}$$

**TABLE 8.6a Example of Table 8.4a: Data of Table 8.5**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = 43{,}687$ |
| $\alpha$-classes (after mean) | 2 | $R(\alpha\|\mu) = 310$ |
| Covariate (pooled within-class regression) | 1 | $R(b\|\mu, \alpha) = 1.5$ |
| Residual error | 3 | SSE = 80.5 |
| Total | 7 | $\text{SST}_{yy} = 44{,}079$ |

from (34). Hence, the results in Table 8.4a become those in Table 8.6a. It can be checked that $R(\mathbf{a}, \mathbf{b})$ of the general case, $R(\mu, \alpha, \beta)$ here, is

$$\begin{aligned} R(\mu, \alpha, b) = R(\mathbf{a}, \mathbf{b}) &= \mathbf{a}^{\circ\prime}\mathbf{X}'\mathbf{y} + \hat{b}\mathbf{Z}'\mathbf{y} \\ &= 71.5(219) + 76.5(156) + 86.5(178) + 0.5(2018) \\ &= 43{,}998.5 \\ &= 43{,}687 + 310 + 1.5 \text{ of Table 8.6a} \\ &= \text{SSM}_{yy} + \text{SSR}_{m,yy} + \text{SSRB of Table 8.4a} \end{aligned}$$

as should be the case.

We can use $F$-statistics available in Table 8.6a for testing hypotheses. From (35),

$$F_{1,3} = \frac{1.5/1}{80.5/3} = 0.06 \text{ tests } H\text{: } b = 0.$$

From (36),

$$F_{2,3} = \frac{310/2}{80.5/3} = 5.8 \text{ tests } H\text{: } \alpha_1 + 3b = \alpha_2 + 3b = \alpha_3 + 5b.$$

Since neither of these $F$-values exceeds the corresponding 5% critical values of 10.13 and 9.55, respectively, we fail to reject both hypotheses.

To calculate Table 8.4b, we get using the basic sums of squares and sums of products,

$$R(b|\mu) = \frac{(2018 - 1975)^2}{(101 - 89.2857)} = 157.8$$

Hence, by subtraction from the sum of the two terms in Table 8.6a,

$$R(\alpha|\mu, b) = 310 + 1.5 - 157.8 = 153.7.$$

Then the results in Table 8.4b become those of Table 8.6b. Since

$$F_{2,3} = \frac{R(\alpha|\mu, b)}{2(80.5)/3} = \frac{153.7(3)}{161} = 2.86$$

is less than the corresponding 5% critical value of 9.55, we fail to reject the hypothesis $H$: $\alpha_1 = \alpha_2 = \alpha_3$ in the covariate model.

**TABLE 8.6b Example of Table 8.4b: Data of Table 8.5**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = 43{,}687$ |
| Covariate (after mean) | 1 | $R(b\|\mu) = 157.8$ |
| $\alpha$-classes (after mean and covariate) | 2 | $R(\alpha\|\mu, b) = 153.7$ |
| Residual error | 3 | SSE = 80.5 |
| Total | 7 | SST = 44,079 |

The following is R output for the above example.

```
> index<-c(74,68,77,76,80,85,93)
> kids<-c(3,4,2,2,4,4,6)
> edu<-c("a","a","a","b","b","c","c")
> result1<-lm(index~kids+edu)
> result2<-lm(index~edu+kids)
> anova(result1)
Analysis of Variance Table

Response: index
          Df Sum Sq Mean Sq F value Pr(>F)
kids       1 157.84 157.841 5.8823 0.09372 .
edu        2 153.66  76.829 2.8632 0.20157
Residuals  3 80.50   26.833
- - -
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(result2)
Analysis of Variance Table

Response: index
          Df Sum Sq Mean Sq F value Pr(>F)
edu        2 310.0 155.000 5.7764 0.0936 .
kids       1   1.5   1.500 0.0559 0.8283
Residuals  3  80.5  26.833
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(result1)
SAS output follows.
```

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **edu** | 3 | 1 2 3 |
| **Number of Observations Read** | | 7 |
| **Number of Observations Used** | | 7 |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: index**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 311.5000000 | 103.8333333 | 3.87 | 0.1479 |
| Error | 3 | 80.5000000 | 26.8333333 | | |
| Corrected Total | 6 | 392.0000000 | | | |
| R-Square | Coeff Var | | Root MSE | index Mean | |
| 0.794643 | 6.557076 | | 5.180090 | 79.00000 | |
| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
| kids | 1 | 157.8414634 | 157.8414634 | 5.88 | 0.0937 |
| edu | 2 | 153.6585366 | 76.8292683 | 2.86 | 0.2016 |
| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
| kids | 1 | 1.5000000 | 1.5000000 | 0.06 | 0.8283 |
| edu | 2 | 153.6585366 | 76.8292683 | 2.86 | 0.2016 |

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| edu | 3 | 1 2 3 |
| Number of Observations Read | | 7 |
| Number of Observations Used | | 7 |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: index**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 311.5000000 | 103.8333333 | 3.87 | 0.1479 |
| Error | 3 | 80.5000000 | 26.8333333 | | |
| Corrected Total | 6 | 392.0000000 | | | |
| R-Square | Coeff Var | | Root MSE | index Mean | |
| 0.794643 | 6.557076 | | 5.180090 | 79.00000 | |
| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
| edu | 2 | 310.0000000 | 155.0000000 | 5.78 | 0.0936 |
| kids | 1 | 1.5000000 | 1.5000000 | 0.06 | 0.8283 |
| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
| edu | 2 | 153.6585366 | 76.8292683 | 2.86 | 0.2016 |
| kids | 1 | 1.5000000 | 1.5000000 | 0.06 | 0.8283 |

Code

```
Data investment;
Input index kids edu;
Cards;
74 3 1
68 4 1
77 2 1
76 2 2
80 4 2
85 4 3
93 6 3
proc glm;
class edu;
model index=kids edu;
proc glm;
class edu;
model index =edu kids;
run;
```

□

**(*iii*) *The Intra-Class Regression Model.*** In (25), we applied the general procedure for covariance analysis to the one-way classification with a solitary covariate and a single regression coefficient $b$. We now show how the general procedure applies when the covariate occurs in the model in some fashion other than the simple case of (25). We consider one alternative (an easy one). In all three cases, $\mathbf{a}^*$ and $R(\mathbf{a})$ are the same for the model (25).

The model based on (25) assumes the same regression slope for all classes. This need not necessarily be the case. One possible alternative is the model

$$y_{ij} = \mu + \alpha_i + b_i z_{ij} + e_{ij} \tag{46}$$

where there is a different regression for each class. We call this an *intra-class regression model.*

The general procedure proceeds quite straightforwardly for this model. Compared to (25), $\mathbf{a}^*$ and $R(\mathbf{a})$ remain the same but $\mathbf{b}$ and $\mathbf{Z}$ are changed. The vector $\mathbf{b}$ is that of the regression slopes and $\mathbf{Z}$ is an $N \times c$ matrix. We have that

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{z}_c \end{bmatrix} = \mathbf{D}\{\mathbf{z}_i\} = \sum_{i=1}^{c} {}^{+}\mathbf{z}_i, \tag{47}$$

for $\mathbf{z}_i$ being the vector of $n_i$ observed $z$'s in the $i$th class.

Applying to each column of Z in (47) the derivation of the corresponding vector of residuals shown in (29) for $\mathbf{y}$, it follows that $\mathbf{R}_z$ of (20) is

$$\mathbf{R}_z = \begin{bmatrix} \mathbf{z}_1 - \bar{z}_{1.}1_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_2 - \bar{z}_{2.}1_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{z}_c - \bar{z}_{c.}1_{n_c} \end{bmatrix} = \sum_{i=1}^{c} {}^{+}(\mathbf{z}_i - \bar{z}_{i.}1_{n_i}). \quad (48)$$

Hence for $\hat{\mathbf{b}}$ of (22), $\mathbf{R}_z'\mathbf{R}_z$ is the diagonal matrix

$$\begin{aligned} \mathbf{R}_z'\mathbf{R}_z &= \mathbf{D}\{(\mathbf{z}_i - \bar{z}_{i.}\mathbf{1}_{n_i})'(\mathbf{z}_i - \bar{z}_{i.}\mathbf{1}_{n_i})\} \\ &= \mathbf{D}\left\{\sum_{j=1}^{n_i} z_{ij}^2 - n_i\bar{z}_{i.}^2\right\} \quad \text{for} \quad i = 1, 2, \ldots, c. \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbf{R}_z'\mathbf{y} &= \{(\mathbf{z}_i - \bar{z}_{i.}\mathbf{1}_{n_i})'\mathbf{y}_i\} \\ &= \left\{\sum_{j=1}^{n_i} y_{ij}z_{ij} - n_i\bar{y}_{i.}\bar{z}_{i.}\right\} \quad \text{for} \quad i = 1, 2, \ldots, c. \end{aligned}$$

Define

$$(\mathrm{SSE}_{zz})_i = \sum_{j=1}^{n_i} z_{ij}^2 - n_i\bar{z}_{i.}^2 \quad \text{and} \quad (\mathrm{SSE}_{yz})_i = \sum_{j=1}^{n_i} y_{ij}z_{ij} - n_i\bar{y}_{i.}\bar{z}_{i.}. \quad (49)$$

Using the two expressions between (48) and (49), we then have

$$\mathbf{R}_z'\mathbf{R}_z = \mathbf{D}\{(\mathrm{SSE}_{zz})_i\} \quad \text{and} \quad \mathbf{R}_z'\mathbf{y} = \mathbf{D}\{(\mathrm{SSE}_{yz})_i\} \quad (50)$$

so that

$$\hat{\mathbf{b}} = (\mathbf{R}_z'\mathbf{R}_z)^{-1}\mathbf{R}_z'\mathbf{y} = \left\{\frac{(\mathrm{SSE}_{yz})_i}{(\mathrm{SSE}_{zz})_i}\right\}.$$

Coordinate-wise, we have

$$\hat{b}_i = \frac{(\mathrm{SSE}_{yz})_i}{(\mathrm{SSE}_{zz})_i}, \quad \text{for} \quad i = 1, 2, \ldots, c. \quad (51)$$

Then with $\mathbf{a}^*$ of (27), we get $\mathbf{a}^\circ$ from (23) as

$$\mathbf{a}^\circ = \begin{bmatrix} \mu^\circ \\ \{\alpha_i^\circ\} \end{bmatrix} = \begin{bmatrix} 0 \\ \{\bar{y}_{i.} - \hat{b}_i\bar{z}_{i.}\} \end{bmatrix} \quad \text{for} \quad i = 1, 2, \ldots, c \quad (52)$$

Thus, from (51), we see that $\hat{b}_i$ is the within-class regression estimator of $y$ on $z$ within the $i$th class and $\alpha_i^\circ$ in (52) is the corresponding intercept estimator for that class. Notice, too, from the definitions in (49) and the result in (51), that the sums of the numerators and denominators of the $\hat{b}_i$ are, respectively, the numerator and denominator of the pooled within-class estimator of (33).

For the model (46), we have

$$R(\mu, \alpha) = R(\mathbf{a}) = \sum_{i=1}^{c} \frac{y_{i\cdot}^2}{n_i} = \mathrm{SSR}_{yy}$$

as before in (28). From (22), using (50) and (51) we have

$$R(\mathbf{b}|\mu, \alpha) = \hat{\mathbf{b}}'\mathbf{R}_z'\mathbf{y} = \sum_{i=1}^{c} \frac{(\mathrm{SSE}_{yz})_i^2}{(\mathrm{SSE}_{zz})_i}. \tag{53}$$

We may use these reductions in the analysis of variance to fit the model (46), along the lines of Table 8.3a. However, it is more instructive to also incorporate Table 8.4a and establish a test of hypothesis $H$: all $b_i$'s equal for the model (46). This is achieved by subtracting $R(b|\mu, \alpha)$ of Table 8.4a from $R(\mathbf{b}|\mu, \alpha)$ of (53). Thus,

$$R(\mathbf{b}|\mu, \alpha) - R(b|\mu, \alpha)$$

is the numerator for testing $H$: all $b_i$'s equal in the model (46). The complete analysis is shown in Table 8.7.

If we estimate $\sigma^2$ by

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{N - 2c}$$

**TABLE 8.7 Analysis of Variance for Fitting the Model $y_{ij} = \mu + \alpha_i + b_i z_{ij} + e_{ij}$ for the One-Way Classification**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = \mathrm{SSM}_{yy}$ |
| $\alpha$-classes (after mean) | $c - 1$ | $R(\alpha\|\mu) = \mathrm{SSR}_{m,yy}$ |
| Covariate (within-class) | $c$ | $R(\mathbf{b}\|\mu, \alpha) = \sum_{i=1}^{c} \frac{(\mathrm{SSE}_{yz})_i^2}{(\mathrm{SSE}_{zz})_i}$ |
| Pooled | 1 | $R(b\|\mu, \alpha) = \frac{(\mathrm{SSE}_{yz})^2}{\mathrm{SSE}_{zz}}$ |
| Difference ($H$: $b_i$'s equal) | $c - 1$ | $R(\mathbf{b}\|\mu, \alpha) - R(b\|\mu, \alpha)$ |
| Residual error | $N - 2c$ | $\mathrm{SSE} = \mathrm{SSE}_{yy} - R(\mathbf{b}\|\mu, \alpha)$ |
| Total | $N$ | $\mathrm{SST}_{yy}$ |

we can use the $F$-statistic

$$F = \frac{R(\mathbf{b}|\mu, \alpha) - R(b|\mu, \alpha))}{(c-1)\hat{\sigma}^2} \tag{54}$$

to test $H$: all $b_i$'s equal. Failure to reject this hypothesis can lead to estimating the pooled $b$ as in (33). The $F$-statistic based on (40) then provides a test, under the assumption of equal $b_i$'s of the hypothesis that the $\alpha_i$'s are equal. The statistic

$$F = \frac{R(b|\mu, \alpha)}{\hat{\sigma}^2} \tag{55}$$

is also available for testing that this pooled $b$ is zero. Of course, using it conditionally in this manner, conditional of (54) not being statistically significant changes the nominal probability level of any critical value used for (55) from that customarily associated with it.

When the hypothesis $H$: all $b_i$'s equal is rejected, one can develop a test of hypothesis $H$: all $\alpha_i$'s equal. However, the interpretation of equal $\alpha$'s and unequal $b$'s, that is, of equal intercept and unequal slopes, is often not easy. It implies a model in the form of a pencil of regression lines through the common intercept. Development of the test is left to the reader. In this case, hypothesis in (17) takes the form

$$H\text{: } \alpha_i + b_i \bar{z}_{i.} \text{ equal for all } i.$$

The $F$-statistic for testing this hypothesis is

$$F = \frac{R(\alpha|\mu)}{(c-1)\hat{\sigma}^2}.$$

**(*iv*) *Continuation of Example 1.***

**Example 2 Estimates for Intra-class Regression Model** We can estimate the within-class regression slopes from (51) using the Table 8.5 data. We obtain

$$\hat{b}_1 = -4.5, \quad \hat{b}_2 = 2 \quad \text{and} \quad \hat{b}_3 = 4.$$

From substitution into (49) and (53), we get

$$R(\mathbf{b}|\mu, \alpha) = \frac{(-9)^2}{2} + \frac{4^2}{2} + \frac{8^2}{2} = 80.5$$

Hence,

$$\text{SSE} = \text{SSE}_{yy} - R(\mathbf{b}|\mu, \alpha) = 82 - 80.5 = 1.5.$$

**TABLE 8.8 Example of Table 7.7: Data of Table 8.5 (See Table 8.6a Also)**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $R(\mu) = 43{,}687$ |
| $\alpha$-classes (after mean) | 2 | $R(\alpha\|\mu) = 310$ |
| Covariate (within class) | 3 | $R(\mathbf{b}\|\mu, \alpha) = 80.5$ |
| Pooled | 1 | $R(b\|\mu, \alpha) = 1.5$ |
| Difference | 2 | Difference = 79 |
| Residual error | 1 | SSE = 1.5 |
| Total | 7 | SST = 44,079 |

Table 8.7 therefore becomes Table 8.8 (based on Table 6.6a). The residual error sum of squares is very small in this example. This is because two of the classes for which the within-class regressions have been estimated have only two sets of observations (see Table 8.5). As a result, the estimation for these two classes is a perfect fit. The only contribution to the residual error is from the one class having three observations. Table 8.5, of course, is not a statistically interesting example. Its sole purpose is to illustrate the derivation of the analysis.

Here are R and SAS outputs showing the sum of squares breakdown for the three covariates.

```
> index<-c(74,68,77,76,80,85,93)
> edu<-c("a","a","a","b","b","c","c")
> kids1<-c(3,4,2,0,0,0,0)
> kids2<-c(0,0,0,2,4,0,0)
> kids3<-c(0,0,0,0,0,4,6)
> result<-lm(index~edu+kids1+kids2+kids3)
> anova(result)
> summary(result)
Analysis of Variance Table

Response: index
          Df Sum Sq Mean Sq F value Pr(>F)
edu        2 310.0 155.0 103.3333 0.06939 .
kids1      1  40.5  40.5  27.0000 0.12104
kids2      1   8.0   8.0   5.3333 0.26015
kids3      1  32.0  32.0  21.3333 0.13574
Residuals  1   1.5   1.5
- - -
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

data investment;
input index edu kids1 kids2 kids3;
cards;
74 1 3 0 0
68 1 4 0 0
77 1 2 0 0
```

```
76 2 0 2 0
80 2 0 4 0
85 3 0 0 4
93 3 0 0 6
proc glm;
class edu;
model index=edu kids1 kids2 kids3;
estimate 'kids1=0' kids1 1;
estimate 'kids2=0' kids2 1;
estimate 'kids3=0' kids3 1;
run;
```

**The SAS System**

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **edu** | 3 | 1 2 3 |
| **Number of Observations Read** | | 7 |
| **Number of Observations Used** | | 7 |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: index**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 390.5000000 | 78.1000000 | 52.07 | 0.1048 |
| **Error** | 1 | 1.5000000 | 1.5000000 | | |
| **Corrected Total** | 6 | 392.0000000 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **index Mean** | |
| 0.996173 | 1.550310 | | 1.224745 | 79.00000 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **edu** | 2 | 310.0000000 | 155.0000000 | 103.33 | 0.0694 |
| **kids1** | 1 | 40.5000000 | 40.5000000 | 27.00 | 0.1210 |
| **kids2** | 1 | 8.0000000 | 8.0000000 | 5.33 | 0.2601 |
| **kids3** | 1 | 32.0000000 | 32.0000000 | 21.33 | 0.1357 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **edu** | 2 | 28.34210526 | 14.17105263 | 9.45 | 0.2242 |
| **kids1** | 1 | 40.50000000 | 40.50000000 | 27.00 | 0.1210 |
| **kids2** | 1 | 8.00000000 | 8.00000000 | 5.33 | 0.2601 |
| **kids3** | 1 | 32.00000000 | 32.00000000 | 21.33 | 0.1357 |
| **Parameter** | **Estimate** | | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **kids1=0** | −4.50000000 | | 0.86602540 | −5.20 | 0.1210 |
| **kids2=0** | 2.00000000 | | 0.86602540 | 2.31 | 0.2601 |
| **kids3=0** | 4.00000000 | | 0.86602540 | 4.62 | 0.1357 |

The estimates of the $b_i$ are given above in the SAS output. □

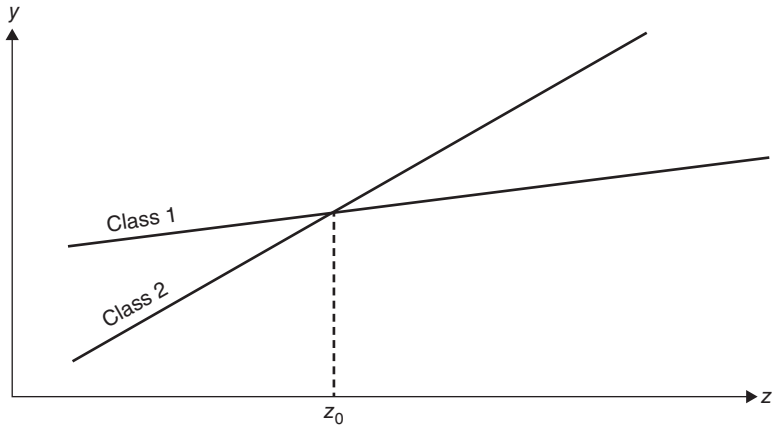



**FIGURE 8.1** Estimated Regression Lines of $y$ on $z$ for Two Classes

**(*v*) *Another Example.***[2] Consider the case of just two classes in a one-way classification. Then $R(\alpha|\mu)$ reduces to $n_1 n_2(\bar{y}_{1.} - \bar{y}_{2.})^2/n$. and tests the hypothesis $H$: $\alpha_1 + b_1\bar{z}_1 = \alpha_2 + b_2\bar{z}_2$. Suppose that the observed means of the two classes are the same, $\bar{y}_{1.} = \bar{y}_{2.}$, or nearly so.

Then $R(\alpha, \mu) = 0$ and we fail to reject the hypothesis. However, we must not draw the conclusion that there is no significant difference between the classes at other values of $z$.

Differences between $\alpha_1 + b_1 z$ and $\alpha_2 + b_2 z$ may be very real for certain values of $z$. Suppose, for example, that the estimated regression lines have the appearance of Figure 8.1. For certain values of $z$ greater than $z_0$, the adjusted value of $y$ for class 2 might be significantly greater than that for class 1. Similarly, for certain values of $z$ less than $z_0$, the mean adjusted $y$-response for class 2 may be significantly less than class 1. A numerical illustration of this is provided in Exercise 3.

### c. The Two-Way Classification (With Interaction)

The purpose of this section is to briefly indicate how to apply the results of the preceding sub-sections 2a and 2b of the present chapter to the two-way classifications (with interaction). We do this in a similar manner to the application for the one-way classification.

The starting point will be $\mathbf{a}^*$ and $R(\mathbf{a})$ for the no-covariate two-way classification (with interaction) model. Recall the discussion of this model in Section 2 of Chapter 7. From equations (55) and (61) of Chapter 7

$$\mathbf{a}^* = \begin{bmatrix} \mathbf{0} \\ \bar{\mathbf{y}} \end{bmatrix} \quad \text{and} \quad R(\mathbf{a}) = \sum_{i=1}^{c} \sum_{j=1}^{b} \frac{y_{ij}^2}{n_{ij}}, \tag{56}$$

[2] S. R. Searle is grateful to E. C. Townsend for bringing this to his notice.

where $\bar{\mathbf{y}}$ is the vector of cell means $\bar{y}_{ij.}$. We also have

$$y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{ij.} \tag{57}$$

as a typical element in the vector of residuals for fitting the no-covariate model. It defines the basis for defining $\mathbf{R}_z$, whose columns are the vectors of residuals that we obtain from the column of $\mathbf{Z}$.

A frequently seen model for covariance in the two-way classification is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + bz_{ijk} + e_{ijk}. \tag{58}$$

Often, we just consider the no-interaction case with $\gamma_{ij}$ omitted. Sometimes the covariate takes the form $b(z_{ij} - \bar{z})$ rather than $bz_{ijk}$. (See, for example, Federer (1955, p. 487) and Steel and Torrie (1960, p. 309).) The form $bz_{ijk}$ seems preferable because then the equation of the model does not involve a sample (i.e., observed) mean. This is appropriate because models should be in terms of population parameters and not observed samples. Moreover, the form $bz_{ijk}$ is more tractable for the general procedure considered earlier, especially if we consider models more complex than (58).

Although (58) is the most commonly occurring model for handling a covariate in the two-way classification, we can also consider other models. The model (58) assumes the same regression slope for all of the cells. The model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + b_i z_{ijk} + e_{ijk} \tag{59}$$

assumes different slopes for each level of the $\alpha$-factor. Likewise, the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + b_j z_{ijk} + e_{ijk} \tag{60}$$

assumes a different slope for each level of the $\beta$-factor. Both of the models

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + (b_i + b_j) z_{ijk} + e_{ijk} \tag{61}$$

and

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + b_{ij} z_{ijk} + e_{ijk} \tag{62}$$

assume different slopes for each $(i, j)$-cell.

We can handle each of the five models (58)–(62) by the general method based on $\mathbf{a}^*$ and $R(\mathbf{a})$ of (56), and on deriving each column of $\mathbf{R}_z$ from the procedure indicated in (57). We determine the exact form of $\mathbf{Z}$ in the general model (6) from the form of the $b$-coefficients in (58)–(62). For example, in (58) $\mathbf{Z}$ is an $N \times 1$ vector, of all the observed $z$'s. In (59) for $c$ levels of the $\alpha$-factor, it is an $N \times c$ matrix of the same form as (48). We can determine the form of $\mathbf{Z}$ for the other models for the $b$-coefficients

(see Exercise 4). We can use the analyses of variance of Tables 8.3a and 8.3b for all of the models (58)–(62).

We can fit the different models in (58)–(62) by using Table 8.4a in the same way that it was used to develop Table 8.7 when fitting $y_{ij} = \mu + \alpha_i + b_i z_{ij} + e_{ij}$ after having fitted $y_{ij} = \mu + \alpha_i + bz_{ij} + e_{ij}$. For each of (58)–(62), we calculate $R(\mathbf{a})$ as in (56). It represents $R(\mu, \alpha, \beta, \gamma)$. We can partition this in either of the two ways indicated in Table 7.8. We derive the hypotheses corresponding to these partitionings, using (17), from the hypotheses tested in the no-covariate model discussed in Sections 2f(ii)–2f(v) of Chapter 7. (In no-interaction analogues of (58)–(62), $R(\mathbf{a})$ of Table 8.3a is $R(\mu, \alpha, \beta)$ of (26) in Chapter 7 and can be partitioned as indicated in Table 7.3.) Details, although lengthy, are quite straightforward. We provide a numerical example in Exercise 4.

Covariance procedures for multiple covariates are simple extensions of the methods for one covariate and follow the general procedures discussed above.

The example below indicates why including a covariate may be important in some analyses.

**Example 3 An Illustration of the Importance of the Covariate for Some Analyses** An experiment was conducted to evaluate the effects of environmental enrichment on intellectual development. The researcher manipulated two levels of an environmental complexity variable (A), and three levels of an age variable (B). Randomly sampled groups of rats were exposed to either $a_1$ or $a_2$ at three ages ($b_1$, $b_2$, and $b_3$, respectively).

As adults, they were tested in a discrimination-learning task (Y). The researcher was concerned that alertness to visual stimulation might be a covariate of influence in the learning task. For this reason, the researcher took a measure (X) of visual attentiveness prior to the treatment. The data below are from page 839 of Winer, Brown and Michels (1991) with kind permission of Mc Graw Hill.

| | Age | | | | | |
|---|---|---|---|---|---|---|
| | $b_1$ | | $b_2$ | | $b_3$ | |
| Complexity | X | Y | X | Y | X | Y |
| $a_1$ | 45 | 100 | 35 | 90 | 55 | 95 |
| | 40 | 85 | 45 | 105 | 45 | 90 |
| | 45 | 100 | 50 | 90 | 45 | 95 |
| | 55 | 110 | 45 | 95 | 35 | 85 |
| | 50 | 105 | 45 | 95 | 45 | 90 |
| $a_2$ | 55 | 105 | 55 | 105 | 50 | 100 |
| | 35 | 100 | 35 | 95 | 35 | 90 |
| | 40 | 100 | 45 | 100 | 30 | 80 |
| | 50 | 115 | 50 | 95 | 55 | 110 |
| | 35 | 90 | 45 | 100 | 40 | 90 |

An SAS output follows.

**The GLM Procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **complex** | 2 | 1 2 |
| **age** | 3 | 1 2 3 |
| **Number of Observations Read** | | 30 |
| **Number of Observations Used** | | 30 |

**The GLM Procedure**
**Dependent Variable: y**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 6 | 1295.775058 | 215.962510 | 7.90 | 0.0001 |
| **Error** | 23 | 628.391608 | 27.321374 | | |
| **Corrected Total** | 29 | 1924.166667 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **y Mean** | |
| 0.673421 | 5.397919 | | 5.226985 | 96.83333 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **complex** | 1 | 67.5000000 | 67.5000000 | 2.47 | 0.1297 |
| **age** | 2 | 361.6666667 | 180.8333333 | 6.62 | 0.0054 |
| **complex∗age** | 2 | 5.0000000 | 2.5000000 | 0.09 | 0.9129 |
| **x** | 1 | 861.6083916 | 861.6083916 | 31.54 | <.0001 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **complex** | 1 | 136.2838655 | 136.2838655 | 4.99 | 0.0355 |
| **age** | 2 | 267.8252313 | 133.9126156 | 4.90 | 0.0169 |
| **complex∗age** | 2 | 12.4182904 | 6.2091452 | 0.23 | 0.7985 |
| **x** | 1 | 861.6083916 | 861.6083916 | 31.54 | <.0001 |

Notice that the complexity is not statistically significant, whereas the age and the covariate are highly significant. Interaction is not significant. Fitting the covariate first, we get

**The GLM procedure**

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **complex** | 2 | 1 2 |
| **age** | 3 | 1 2 3 |
| **Number of Observations Read** | | 30 |
| **Number of Observations Used** | | 30 |

**The GLM Procedure**

**Dependent Variable: y**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 6 | 1295.775058 | 215.962510 | 7.90 | 0.0001 |
| **Error** | 23 | 628.391608 | 27.321374 | | |
| **Corrected Total** | 29 | 1924.166667 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **y Mean** | |
| 0.673421 | 5.397919 | | 5.226985 | 96.83333 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **x** | 1 | 875.2924217 | 875.2924217 | 32.04 | <.0001 |
| **complex** | 1 | 138.3768433 | 138.3768433 | 5.06 | 0.0343 |
| **age** | 2 | 269.6875029 | 134.8437514 | 4.94 | 0.0165 |
| **complex*age** | 2 | 12.4182904 | 6.2091452 | 0.23 | 0.7985 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **x** | 1 | 861.6083916 | 861.6083916 | 31.54 | <.0001 |
| **complex** | 1 | 136.2838655 | 136.2838655 | 4.99 | 0.0355 |
| **age** | 2 | 267.8252313 | 133.9126156 | 4.90 | 0.0169 |
| **complex*age** | 2 | 12.4182904 | 6.2091452 | 0.23 | 0.7985 |

Given the covariate, we see that both complexity and age are statistically significant. Thus, the results of the analysis are affected by the covariate. □

## 3. DATA HAVING ALL CELLS FILLED

Analysis of unbalanced data is more difficult than that of balanced data, for the very reason that such data are unbalanced. Often interpretation of the analyses is more difficult. Sometimes, if the unbalanced data are not too far from being balanced, the difficulties may be avoided. In such cases, it is sometimes possible to make minor modifications in the data so as to be able to use a balanced data analysis. The decision whether or not to do this depends on the answer to the following question. When are unbalanced data "not too far removed" from being balanced data? It is highly unlikely that a satisfactory answer to this question can be given. Nevertheless, the advantages of using a balanced data analysis are so great that one would like to use then whenever feasible. Balanced data analyses are much more easily carried out and interpreted in comparison with analogous unbalanced data analyses.

The disadvantage of modifying unbalanced data so as to be able to use a balanced data analysis is that in doing so it introduces a measure of approximation into the analysis. The degree of the approximation depends on the extent to which the unbalanced data have been modified to permit the balanced analysis. However, with the advantages of balanced data analysis being so attractive they may, on occasion outweigh the disadvantage of some degree of approximation if this approximation may be deemed small. We outline instances in which this might be so, below. To simplify presentation, we use examples for the two-way crossed classification.

**TABLE 8.9 $n_{ij}$-Values**

| | | |
|---|---|---|
| 6 | 6 | 5 |
| 6 | 6 | 6 |

### a. Estimating Missing Observations

If all the $n_{ij}$'s except a few are the same, it is often reasonable to estimate missing observations. For example, suppose with two rows and three columns, the number of observations are as shown in Table 8.9. Data of this nature often arise from what set out to be a planned experiment (in Table 8.9 of six observations per cell) and ended up with a few observations missing. Such data are unbalanced, but so slightly as to render the temptation of making them balanced irresistible. We can accomplish this by estimating the missing observations. In this case, we need to estimate one observation for the cell in the first row and the third column. One way to do this is to suppose that $u$, say, represents the missing observation and choose $u$ so as to minimize the residual sum of squares. If there were no missing values, $n_{13}$ would have been six instead of five. For an interactions model (see equation (61) of Chapter 7) would have then been

$$\text{SSE} = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{6} y_{ijk}^2 - \frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij.}^2.$$

For the missing value data with $u$ representing the missing value, we would have for the residual,

$$\text{SSE} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6} y_{ijk}^2 + \sum_{k=1}^{6} y_{23k}^2 + \sum_{k=1}^{5} y_{13k}^2 + u^2 - \frac{1}{6}\left[\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij.}^2 + y_{23.}^2 + (y_{13.} + u)^2\right].$$

In order to minimize this quantity, we obtain $\partial(\text{SSE})/\partial \text{u}$, set it equal to zero, and solve for $u$ in terms of the observations in SSE.

$$\frac{\partial(\text{SSE})}{\partial u} = 2u - \frac{1}{3}(y_{13.} + u) = 0.$$

The solution to the above equation is

$$u = \frac{1}{5}\sum_{k=1}^{5} y_{13k} = \bar{y}_{13.}.$$

The second derivative of SSE is $5/3 > 0$ so that SSE is indeed minimized. As a result, the missing observation in the (1, 3) cell is estimated by the mean of the observations that are there.

Of course, the form of the results arising from such a process depends on the model used. This determines the residual sum of squares. Had the model for the data of Table 8.9 been that of no interaction, the error sum of squares would have been (see Section 1 of Chapter 7)

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6} y_{ijk}^2 + \sum_{k=1}^{6} y_{23k}^2 + u^2 - \frac{1}{18}\left[(y_{1..}+u)^2 + y_{2..}^2\right] - \frac{1}{12}\left[y_{.1}^2 + y_{.2}^2 + (y_{.3.}+u)^2\right] + \frac{1}{36}(y_{...}+u)^2.$$

Minimization with respect to $u$ of the analogous quantity for the general case of $a$ rows, $b$ columns, and $n$ observations per cell in all cells except one the $(i, j)$th cell to

$$u_{ij} = \frac{ax_{i..} + bx_{.j.} - x_{...}}{ab(n-1) + (a-1)(b-1)}. \tag{63}$$

Equation (63) is equivalent to the result given by Federer (1955, p, 134, equation V-52) for $n = 1$. Federer also gives results for more than one missing observation when $n = 1$. (These are the procedures referred to at the beginning of Section 1 of Chapter 7.)

Bartlett (1937) presents a generalization of the above procedure that depends on a covariance technique. For the model $\mathbf{y} = \mathbf{Xa} + \mathbf{Zb} + \mathbf{e}$, this involves doing the following:

(i) in $\mathbf{y}$, include each missing observation as an observation of zero;
(ii) in $\mathbf{b}$, include, negatively, a parameter for each missing observation;
(iii) in $\mathbf{Z}$, have one column for each parameter mentioned in (ii), all entries being zero except for a single unity corresponding to the $y$-value of zero specified in (i).

One will find that the normal equations of this covariate model are satisfied by the estimated missing observations derived by minimizing residual sums of squares as derived earlier. For example, for the data of Table 8.9 (without row-by-column interactions) has the following normal equations:

$$\begin{bmatrix} 36 & 18 & 18 & 12 & 12 & 12 & 1 \\ 18 & 18 & 0 & 6 & 6 & 6 & 1 \\ 18 & 0 & 18 & 6 & 6 & 6 & 0 \\ 12 & 6 & 6 & 12 & 0 & 0 & 0 \\ 12 & 6 & 6 & 0 & 12 & 0 & 0 \\ 12 & 6 & 6 & 0 & 0 & 12 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu^\circ \\ \alpha_1^\circ \\ \alpha_2^\circ \\ \beta_1^\circ \\ \beta_2^\circ \\ \beta_3^\circ \\ -u^\circ \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1..} \\ y_{2..} \\ y_{.1.} \\ y_{.2.} \\ y_{.3.} \\ 0 \end{bmatrix} \tag{64}$$

The reader may show in Exercise 5 that the appropriate form of (63) is a solution to (64). This procedure leads to the same results as minimizing residual sums of squares. However, it is often computationally much easier because it can be applied directly by means of analysis of covariance procedures (see Section 2).

Rao and Toutenburg (1999) call attention to a procedure suggested by Yates (1933).

Suppose we have what would be a balanced model with $t$ missing observations. He suggests that we reorganize the data matrix according to

$$\begin{bmatrix} y_{obs} \\ y_{mis} \end{bmatrix} = \begin{bmatrix} X_{ub} \\ X_* \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_{ub} \\ \varepsilon_* \end{bmatrix}$$

The $X_{ub}$ would represent the levels of the missing values for the unbalanced model we would use without the missing observations. The $X_*$ would correspond to the entries in the design matrix for the balanced model that would correspond to the missing $y$ values. Now find the least-square estimator for the unbalanced model using any generalized inverse of $\mathbf{X}'_{ub}\mathbf{X}_{ub}$. Then,

$$\mathbf{b}_{ub} = (\mathbf{X}'_{ub}\mathbf{X}_{ub})^{-}\mathbf{X}'_{ub}y_{obs}.$$

We may now replace the missing value by

$$\hat{\mathbf{y}}_{mis} = \mathbf{X}_*\mathbf{b}_{ub}.$$

The reader may establish the equivalence of this method to the one considered above.

Rao and Toutenburg (1999) also suggest using a shrinkage estimator of the James–Stein type (see Gruber (1998)). It takes the form

$$\hat{y}_{mis} = \left(1 - \frac{k\hat{\sigma}^2_{ub}}{(N_{ub} - m + 2)\mathbf{b}'_{ub}\mathbf{X}'_{ub}\mathbf{X}_{ub}\mathbf{b}_{ub}}\right)\mathbf{X}_*\mathbf{b}_{ub}.$$

We use estimates of missing observations just as if they were data. There is, however, one change that must be made in the balanced data analysis of the combined data (observed and missing). The degrees of freedom for the residual error sum of squares are calculated as for balanced data and then reduced for the number of missing observations that have been estimated. Thus in an interaction analysis of data like those in Table 8.9, the residual sum of squares for six observations in every cell would be 6(5) = 30. However, with one estimated missing observation, it is reduced to 29.

**Example 4 A Word of Caution** Sometimes the method of minimizing the residual sum of squares can lead to replacement values that are not in keeping with the physical nature of the problem being solved. We give an example of this.

Consider the data in Table 7.1. We denote the missing observation for pan make A and brand of stove Y by $u$, for pan make B and stove brand Y by $v$, and for pan make

B and brand of stove Z by $w$. Now consider minimizing the residual sum of squares.

$$r(u,v,w) = 1728 + u^2 + v^2 + w^2 + \frac{1}{12}(108 + u + v + w)^2 - \frac{1}{3}[3645 + (9 + u + v)^2 + (18 + w)^2] - \frac{1}{4}[4356 + (27 + u)^2 + (15 + v + w)^2]$$

After differentiation and algebraic simplification, we get the equations

$$\frac{1}{6}(-9 + 6u - 3v + w) = 0$$
$$\frac{9}{2} - \frac{u}{2} + v - \frac{w}{3} = 0$$
$$\frac{1}{6}(-9 + u - 2v + 6w) = 0.$$

The solution to this system of equations is $u = -1$, $v = -5$, and $w = 0$. However, the data values are the number of seconds beyond three minutes taken to boil two quarts of water. A negative value would indicate that the water boiled in less than three minutes. Observe that $v = -5$ does not fit the physical nature of the problem if we assume that we started at the same temperature for each pan make and stove brand. □

### b. Setting Data Aside

If the numbers of observations in the sub-most cells differ from each other by only a few, it might not be unreasonable to randomly set aside data from appropriate cells in order to reduce all cells to having the same number of observations in each. Data so reduced can then readily be analyzed as balanced data. For example, in data having the $n_{ij}$ values of Table 8.10, it might be reasonable to randomly set aside observations in order to reduce each cell to 11 observations. This method has some disadvantages. There is inevitable indecisiveness implicit in the suggestion of doing this only when the $n_{ij}$ differ "by only a few." It begs the question "What is a few?" There is no clear-cut answer to this question. All one can say is that the method might be tolerable for $n_{ij}$-values like those in Table 8.10, but not for some like those in Table 8.11. To employ this method for data in Table 8.11, too many data points would have to be set aside. Of course, we have the argument that we should never set any data aside. That is so, except that all good rules have their exceptions. If we believe that balanced data analyses are preferred over those for non-balanced data, it appears to us that randomly setting aside data in cases having $n_{ij}$ like those of Table 8.10 is probably

**TABLE 8.10 $n_{ij}$-Values**

| | | |
|---|---|---|
| 14 | 11 | 13 |
| 11 | 13 | 15 |

**TABLE 8.11 $n_{ij}$-Values**

| | | |
|---|---|---|
| 10 | 17 | 21 |
| 19 | 22 | 9 |

not unreasonable—especially if the within-cell variation is small. Even though we cannot give a clear-cut definition of when to do this and when not to, there will surely be occasions when doing so seems reasonably safe.

At least on these occasions, it would seem to be an acceptable procedure. After all, for the person whose data they are, the case of a balanced data analysis is certainly worthwhile.

This method does not involve discarding data—nor is it described as such—only setting it aside. After setting the data aside and making the balanced data analysis, we can return the data and repeat the process. We can make a different random selection of the data, set it aside and do another balanced analysis. It will of course not be statistically independent of the first analysis. If the conclusions stemming from it are different from those of the first analysis, the result is confusion and not enlightenment. If, when we do further analysis and get additionally different conclusions, we compound the confusion. This confusion might not arise very often for cases where only "a few" observations are set aside and within cell variability is small. Indeed, if such confusion does arise one might suspect that some of the set aside observations might be outliers and be treated as such. Indeed, outliers should probably be set-aside in the first place. Nevertheless, this method should be used with caution. In the worst-case scenario, one can always retreat to the unbalanced data analysis. Perhaps, after a lot of computer simulation, one might be able to put forth some "rules of thumb" as to when this method might be appropriate.

### c. Analysis of Means

**(*i*) *Unweighted Means Analysis.*** When all sub-most cells are filled, an easily calculated analysis is to treat the means of those cells as observations and subject them to a balanced data analysis. This procedure is due to Yates (1934). Of course, this is only an approximate analysis. As usual, the degree of the approximation depends on the extent to which the unbalanced data are not balanced.

The calculations for this analysis are straightforward. The method is known as the unweighted means analysis and proceeds as follows:

Suppose the model for $y_{ijk}$ is as in equation (51) of Section 2a of Chapter 7,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}.$$

For each cell, calculate the mean

$$x_{ij} = \bar{y}_{ij.} = \sum_{k=1}^{n_{ij}} \frac{y_{ijk}}{n_{ij}}.$$

Table 8.12 shows the unweighted means analysis.

**TABLE 8.12 Unweighted Means Analysis for a Two-Way Crossed Classification**

| Source of Variation | d.f. | Sum of Squares | Mean Square |
|---|---|---|---|
| Rows | $a-1$ | $\text{SSA}_u = b\sum_{i=1}^{a}(\bar{x}_{i.}-\bar{x}_{..})^2$ | $\text{MSA}_u$ |
| Columns | $b-1$ | $\text{SSB}_u = a\sum_{j=1}^{b}(\bar{x}_{.j}-\bar{x}_{..})^2$ | $\text{MSB}_u$ |
| Interaction | $(a-1)(b-1)$ | $\text{SSAB}_u = \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$ | $\text{MSAB}_u$ |
| Residual Error | $N-ab$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$ | MSE |

Several facets of Table 8.12 are worth noting.

1. The means of the $x_{ij}$'s are calculated in the usual manner, for example, $\bar{x}_{i.} = \sum_{j=1}^{b} x_{ij}/b$.
2. The residual error sum of squares, SSE, is exactly as calculated in the model for $y_{ijk}$ of Section 2 of Chapter 7.
3. The sum of squares does not add up to SST $= \sum y^2$. The first three add to $\text{SSA}_u$, and $\text{SSB}_u$, and $\text{SSAB}_u$ add to $\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}^2 - x_{..}^2/ab$, but all four do not add to SST.
4. The sums of squares SSA$u$ and SSB$u$ do not have $\chi^2$-distributions nor are they independent of SSE.

Expected values of the mean squares are as follows.

$$\begin{aligned}
E(\text{MSA}_u) &= \frac{b}{a-1}\sum_{i=1}^{a}[\alpha_i + \bar{\gamma}_{i.} - (\bar{\alpha}_{.} + \bar{\gamma}_{..})]^2 + n_h\sigma_e^2 \\
E(\text{MSB}_u) &= \frac{a}{b-1}\sum_{i=1}^{a}[\beta_j + \bar{\gamma}_{.j} - (\bar{\beta}_{.} + \bar{\gamma}_{..})]^2 + n_h\sigma_e^2 \\
E(\text{MSAB}_u) &= \frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\gamma_{ij} - \bar{\gamma}_{i.} - \bar{\gamma}_{.j} + \bar{\gamma}_{..})^2 + n_h\sigma_e^2 \\
E(\text{MSE}) &= \sigma_e^2
\end{aligned} \tag{65}$$

with

$$\frac{1}{n_h} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}},$$

$n_h$ being the harmonic mean of all ab $n_{ij}$'s.

**TABLE 8.13 $n_{ij}$-Values**

| | | |
|---|---|---|
| 192 | 250 | 175 |
| 320 | 168 | 270 |

Since the mean squares of (65) do not have $\chi^2$-distributions, their ratios do not provide $F$-statistics for testing hypotheses. However, Gosslee and Lucas (1965) suggest that they provide reasonably satisfactory $F$-statistics using amended degrees of freedom for the numerator mean squares. For example, the numerator degrees of freedom suggested for $\text{MSA}_u/\text{MSE}$ is

$$f'_a = \frac{(a-1)^2\left(\sum_{i=1}^{a} 1/h_{i.}\right)^2}{\left(\sum_{i=1}^{a} 1/h_{i.}\right)^2 + (a-2)\sum_{i=1}^{a} 1/h_{i.}^2}, \tag{66}$$

where

$$h_{i.} = \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n_{ij}}$$

with $1/h_{i.}$ being the harmonic mean of the $n_{ij}$'s of the cells of the $i$th row. The origin of (66) in Goslee and Lucas (1965) is that of equating the first two moments of $\text{MSA}_u$ to the first two moments of a $\chi^2$-distribution, in the manner of Section 4i of Chapter 2. Although these amended degrees of freedom modify $\text{MSA}_u/\text{MSE}$ to be an approximate $F$-statistic, we see from (65) that the hypothesis it tests is equality of $\alpha_i + \bar{\gamma}_{i.}$ for all $i$. Alternatively, and indeed very reasonably, we can interpret the test as testing equality of the row effects in the presence of the average interaction effects. This hypothesis may often be of interest.

The question attaching to any approximate analysis suggested as a substitute for exact unbalanced analysis remains: when can the unweighted means analysis be used? As usual there is no decisive answer (apart from requiring trivially that $n_{ij} > 0$). Since the unweighted means analysis uses cell means as if they were observations with uniform sampling error, a criterion for using the analysis is to require that these sampling errors be approximately the same. This demands that since the sampling error of a cell mean is proportional to $1/\sqrt{n_{ij}}$ that the values of $1/\sqrt{n_{ij}}$ are approximately equal. What "equal" in this context means is necessarily vague. For example, the values of $1/\sqrt{n_{ij}}$ are approximately equal for the cells of Table 8.11 and for those of Table 8.13, but not for Table 8.14. Unweighted means

**TABLE 8.14 $n_{ij}$-Values**

| | | |
|---|---|---|
| 10 | 17 | 200 |
| 130 | 22 | 9 |

**TABLE 8.15 An Example of Two Rows and Three Columns**

| Column | | | | |
|---|---|---|---|---|
| Row | 1 | 2 | 3 | Total |
| 1 | 7 | 2 | 3 | |
| | 11 | 4 | 9 | |
| | | 6 | | |
| | – | – | – | |
| | 18(2)9 | 12(3)4 | 12(2)6 | 42(7)6 |
| 2 | 11 | 15 | 38 | |
| | 14 | 16 | 46 | |
| | 17 | 19 | | |
| | | 22 | | |
| | – | – | – | |
| | 42(3)14 | 72(4)18 | 84(2)42 | 198(9)22 |
| Total | 60(5)12 | 84(7)12 | | 240(16)15 |

analyses would therefore seem appropriate for data having the values of Table 8.11 or 8.13 but not for Table 8.14. In Table 8.14, we observe that $1/\sqrt{9}$ is more than four times as large as $1/\sqrt{200}$. Maybe a ratio of 2:1 could be tolerated in the values of $1/\sqrt{n_{ij}}$, for using an unweighted data analysis, but probably not a ratio as large as 4:1. The appropriate analysis for Table 8.14 is the unbalanced data analysis.

**(*ii*) *Example.***

**Example 5 Numerical Illustration of Unweighted Analysis of Means** Suppose data for two rows and three columns are as shown in Table 8.15. The layout of data follows the same style as Table 7.6. Each triplet of numbers represents a total of observations, the number of observations in that total in parenthesis, and the corresponding mean.

The unweighted analysis of means of this data is based on the cell means, summarized in Table 8.16. Fitting the model

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

**TABLE 8.16 Cell Means of Table 8.15**

| | Column | | | Total |
|---|---|---|---|---|
| Row | 1 | 2 | 3 | |
| 1 | 9 | 4 | 6 | 19 |
| 2 | 14 | 18 | 42 | 74 |
| Total | 23 | 22 | 48 | 93 |

**TABLE 8.17 Example of Table 8.12: Unweighted Means Analysis of Data of Table 8.15**

| Source of Variation | d.f | Sum of Squares |
|---|---|---|
| Rows | 1 | $SSA_u = 1945.667 - 1441.5 = 504.167$ |
| Columns | 2 | $SSB_u = 1658.5 - 1441.5 = 217$ |
| Interaction | 2 | $SSAB_u = 2417 - 1945.667 - 1658.5 + 1441.6 = 254.333$ |
| Residual error | 10 | SSE = 114 |

to the values of Table 8.16 gives

$$R(\mu) = \frac{93^2}{6} = 1441.5$$
$$R(\mu, \alpha) = \frac{19^2 + 74^2}{3} = 1945.667$$
$$R(\mu, \beta) = \frac{23^2 + 22^2 + 48^2}{2} = 1658.6$$

and

$$R(\mu, \alpha, \beta) = 9^2 + 4^2 + \cdots 42^2 = 2417.$$

From these, we calculate the first three terms of Table 8.12 as shown in Table 8.17. The last term, SSE comes directly from the data of Table 8.15 as

$$\text{SSE} = \left(7^2 + 11^2 - \frac{18^2}{2}\right) + \cdots + \left(38^2 + 46^2 - \frac{84^2}{3}\right) = 114,$$

the sum of the within-cell sum of squares.

We can calculate $F$-statistics in the usual fashion. Observe that equation (66) simplifies to unity when $a = 2$. Thus, by (66) the amended degrees of freedom for $\text{MSA}_u/\text{MSE}$ are $f'_a = 1$. To illustrate the calculation of (66), we derive the comparable value of $f'_b$ as follows. We have

$$\frac{1}{h_{.1}} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{3}\right) = \frac{5}{12}, \frac{1}{h_{.2}} = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{4}\right) = \frac{7}{24}, \text{ and } \frac{1}{h_{.3}} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{1}{2}.$$

Then,

$$\sum_{j=1}^{3} \frac{1}{h_{.j}^2} = \frac{100 + 49 + 144}{24^2} = \frac{293}{24^2}$$

and

$$\sum_{j=1}^{3} \frac{1}{h_{.j}} = \frac{(10 + 7 + 12)}{24} = \frac{29}{24}.$$

By substitution,

$$f'_b = \frac{(3-1)^2(29/24)^2}{(29/24)^2 + (3-2)3(293/24^2)} = \frac{3364}{1720} = 1.96.$$

□

**(*iii*) *Weighted Squares of Means.*** Yates (1934) also suggested an alternative analysis of means known as the weighted squares of means. This is due to Yates (1934). An important advantage of the weighted analysis is that it provides mean squares that *do* have $\chi^2$-distributions. Hence, $F$-statistics will be available for hypothesis testing.

The analysis is based on sums of squares of the means $x_{ij} = \bar{y}_{ij}$ defined in 3c(i). In this analysis, we weight the terms in those sums of squares in inverse proportion to the variance of the term concerned. Thus in place of

$$\mathrm{SSA}_u = b \sum_{i=1}^{a} (\bar{x}_{i.} - \bar{x}_{..})^2$$

of Table 8.12, we use

$$\mathrm{SSA}_w = \sum_{i=1}^{a} w_i (\bar{x}_{i.} - \bar{x}_{[1]})^2,$$

where $w_i$ is $\sigma^2 / v(\bar{x}_{i.})$ and $\bar{x}_{[1]}$ is the weighted mean of the $\bar{x}_i$'s weighted by the $w_i$'s. See Table 8.18 for details. Like the sum of squares in Table 8.12 the sums of squares in Table 8.18 do not add up to $\sum y^2$. However, when the sums of squares are divided by $\sigma^2$ they do have $\chi^2$-distributions. As a result the $F$-statistics $\mathrm{MSA_w/MSE}$, $\mathrm{MSB_w/MSE}$ and $\mathrm{MSAB_w/MSE}$ do provide exact tests of the hypotheses concerning

**TABLE 8.18 Weighted Means Analysis for a Two-Way Crossed Classification**

| Source of Variation | d.f. | Sum of Squares[a] | Mean Square |
|---|---|---|---|
| Rows | $a-1$ | $\mathrm{SSA}_w = \sum_{i=1}^{a} w_i (\bar{x}_{i.} - \bar{x}_{[1]})^2$ | $\mathrm{MSA}_w$ |
| Columns | $b-1$ | $\mathrm{SSB}_w = \sum_{j=1}^{b} v_j (\bar{x}_{.j} - \bar{x}_{[2]})^2$ | $\mathrm{MSB}_w$ |
| Interaction | $(a-1)(b-1)$ | $\mathrm{SSAB}_w = \mathrm{SSAB}_u$ of Table 8.12 $= \sum_{i=1}^{a} \sum_{j=1}^{b} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$ | $\mathrm{MSAB}_w = \mathrm{MSAB}_u$ |
| Residual error | $N-ab$ | $\mathrm{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2$ | MSE |

[a] $w_i = \left( \frac{1}{b^2} \sum_{j=1}^{b} \frac{1}{n_{ij}} \right)^{-1}$ and $\bar{x}_{[1]} = \frac{\sum_{i=1}^{a} w_i \bar{x}_{i.}}{\sum_{i=1}^{a} w_i}$. $v_j = \left( \frac{1}{a^2} \sum_{i=1}^{a} \frac{1}{n_{ij}} \right)^{-1}$ and $\bar{x}_{[2]} = \frac{\sum_{j=1}^{b} v_j \bar{x}_{.j}}{\sum_{j=1}^{b} v_j}$.

the $\alpha$'s, $\beta$'s and $\gamma$'s. We ascertain the exact form of the hypotheses by considering expected values of the mean squares. They are

$$E(\mathrm{MSA}_w) = \frac{1}{a-1}\sum_{i=1}^{a} w_i \left[\alpha_i + \bar{\gamma}_{i.} - \frac{\sum_{i=1}^{a} w_i(\alpha_i + \bar{\gamma}_{i.})}{\sum_{i=1}^{a} w_i}\right]^2 + \sigma_e^2 \tag{67a}$$

and

$$E(\mathrm{MSB}_w) = \frac{1}{b-1}\sum_{j=1}^{b} v_j \left[\beta_j + \bar{\gamma}_{.j} - \frac{\sum_{j=1}^{b} v_j(\beta_j + \bar{\gamma}_{.j})}{\sum_{j=1}^{b} v_j}\right]^2 + \sigma_e^2. \tag{67b}$$

Hence,

$$F = \mathrm{MSA}w/\mathrm{MSE}$$

tests the hypothesis

$$H\text{: } (\alpha_i + \bar{\gamma}_{i.}) \text{ all equal.} \tag{68}$$

As was the case for the unweighted analysis of means [Table 8.12] and the expected values in (65) here, the hypothesis (68) involves the $\bar{\gamma}_i$'s. If, as a restriction on the model, we assume that $\bar{\gamma}_{i.} = 0$ for all $i$, the hypothesis is then one of testing the equality of the $\alpha_i$'s where the weights are the $w_i$'s. Alternatively, without any restriction, the hypothesis is that of testing equality of the row effects in the presence of the average interaction effects. The important difference from the un-weighted analysis is, though, that the $F$-statistics of Table 8.18 have exact $F$-statistics whereas those of Table 8.12 have only approximate $F$-distributions.

We shall return to Tables 8.12 and 8.18 when we discuss variance components in Chapter 10.

**(*iv*) *Continuation of Example.***

**Example 6 Calculation of Table 8.18 for the Data of Tables 8.14 and 8.16** We have that the $w$'s are

$$w_1 = \left[\frac{1}{9}\left(\frac{1}{2} + \frac{1}{3} + \frac{1}{2}\right)\right]^{-1} = \frac{27}{4} \quad \text{and} \quad w_2 = \left[\frac{1}{9}\left(\frac{1}{3} + \frac{1}{4} + \frac{1}{2}\right)\right]^{-1} = \frac{108}{13}.$$

The $v$'s are

$$v_1 = \left[\frac{1}{4}\left(\frac{1}{2}+\frac{1}{3}\right)\right]^{-1} = \frac{24}{5},\ v_2 = \left[\frac{1}{4}\left(\frac{1}{3}+\frac{1}{4}\right)\right]^{-1} = \frac{48}{7}$$
$$\text{and}\quad v_3 = \left[\frac{1}{4}\left(\frac{1}{2}+\frac{1}{2}\right)\right]^{-1} = 4.$$

Now,

$$\bar{x}_{[1]} = \frac{\frac{27}{4}\left(\frac{19}{3}\right)+\frac{108}{13}\left(\frac{74}{3}\right)}{\frac{27}{4}+\frac{108}{13}} = \frac{477}{29}$$

and thus,

$$\text{SSA}_w = \frac{27}{4}\left(\frac{19}{3}-\frac{477}{29}\right)^2 + \frac{108}{13}\left(\frac{74}{3}-\frac{477}{29}\right)^2 = 1251.72.$$

Furthermore,

$$\bar{x}_{[2]} = \frac{\frac{24}{5}\left(\frac{23}{2}\right)+\frac{48}{7}\left(\frac{22}{2}\right)+4\left(\frac{48}{2}\right)}{\frac{24}{5}+\frac{48}{7}+4} = \frac{1983}{7}$$

and

$$\text{SSB}_w = \frac{24}{5}\left(\frac{23}{2}-\frac{1{,}983}{137}\right)^2 + \frac{48}{7}\left(\frac{22}{2}-\frac{1{,}983}{137}\right)^2 + 4\left(\frac{48}{2}-\frac{1{,}983}{137}\right)^2$$
$$= \frac{66{,}882}{137} = 488.19.$$

Table 8.18 therefore becomes as shown in Table 8.19.

**TABLE 8.19 Example of Table 8.18: Weighted Squares of Means Analysis of Data in Table 8.15**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Rows | 1 | $\text{SSA}_w = 1{,}251.72$ |
| Columns | 2 | $\text{SSB}_w = 488.19$ |
| Interaction | 2 | $\text{SSAB}_w = 254.33 = \text{SSAB}_u$ of Table 8.17 |
| Residual error | 10 | SSE = 114 |

□

**TABLE 8.20 $n_{ij}$-Values**

| | | | | | | |
|---|---|---|---|---|---|---|
| (A) | 27 | 32 | 0 | 3 | 1 | (B) |
| | 11 | 12 | 2 | 0 | 2 | |
| (C) | 1 | 0 | 27 | 16 | 24 | (D) |
| | 0 | 8 | 15 | 21 | 22 | |

### d. Separate Analyses

Suppose data had the $n_{ij}$ values shown in Table 8.20. For purposes of discussion, dashed lines divide the data into four sets A, B, C, and D. The only appropriate way of analyzing the complete set of data represented by the $n_{ij}$-values of Table 8.20 would be to use unbalanced data analysis. This is because of the empty cells and widely disparate values of the non-zero $n_{ij}$'s. Such an analysis, using the interaction model of Section 2 of Chapter 7, would provide no testable hypothesis concerning row (or column) effects unencumbered by interactions. Keeping this in mind, observe that in the four cells labeled A, and the six labeled D, all cells are filled. Moreover, in B and C, there are few data and several empty cells. This prompts the suggestion of making two separate analyses, one of the cells A and one of the cells D using an analysis of means in both cases. In analyzing A, comparison between rows 1 and 2 and columns 1 and 2 can be made. Likewise, from analyzing D, comparisons among rows 3 and 4 and columns 3, 4, and 5 could be made. Of course, comparisons that cut across these groups of rows and columns are precluded by such an analysis. However, then the only alternative is an unbalanced data analysis that provides no satisfactory information on such comparisons anyway in the interaction model. Therefore, it would seem just analyzing A and D would cause little to be lost.

When data of the nature alluded to in Table 8.20 occur, one might immediately question the process by which $n_{ij}$-values of such disparate sizes and groupings have arisen. Be that as it may, in analyzing large-scale survey-size data such as are discussed in Section 1, the suggestion has sometimes been made of analyzing just the all-cells-filled subsets of cells that occur throughout the data. Although such a suggestion might be open to criticism, it might not be unreasonable in a small situation such as that envisaged in Table 8.20—should it ever arise. It amounts to analyzing sets of data that are what might be called "*weakly connected*." In Table 4.20, cells labeled B and C do have data in them, but very small amounts compared to A and D. If B and C did not contain any data at all then the sets A and D would be disconnected sets of data and they would *have* to be analyzed separately. As it is, analyzing A and D separately and ignoring B and C would be easy both to compute and interpret. For these reasons, it may be preferable to analyzing the complete data as one analysis.

## 4. EXERCISES

**1** **(a)** Use equation (22) to confirm (42).

**(b)** Write down the normal equations for the data of Table 8.5 using (25) as the model.

**(c)** Derive the solution given in (42) and (44).

**TABLE 8.21 Two Groups of Students**

| Totals | Received Lecture | Received Programmed Text |
|---|---|---|
| $n$ | 31 | 31 |
| $\sum x_1$ | 2139 | 2149 |
| $\sum x_2$ | 3100 | 3100 |
| $\sum x_1^2$ | 148,601 | 157,655 |
| $\sum x_2^2$ | 318,990 | 319,920 |
| $\sum x_1 x_2$ | 216,910 | 224,070 |

**2** For the data of Table 8.5 fit each of the following models and calculate the analyses of variance of Tables 8.4a and 8.4b. Suggest appropriate hypotheses and test them.

**(a)** The covariate affects $y$ linearly, in the same manner for all high school graduates as it does college graduates, but differently for those who did not complete high school.

**(b)** The covariate affects $y$ in both a linear and quadratic manner, the same for everyone.

**3** Townsend (1969) gives data about an experiment designed to determine if the usual lecture-type presentation could be replaced by a programmed text (See Table 8.21.). A class of 62 students was randomly divided into two groups, with one group receiving the usual lectures while the other was given a programmed textbook for independent study. At the end of the semester, both groups were given the same examination. In addition to the final exam score ($x_1$), a measurement of IQ ($x_2$) was recorded for each student. (Other educational studies indicate that the performance may be linearly related to IQ.) Using the basic calculations shown in Table 8.21 carry out a covariance analysis testing any hypothesis you think suitable.

**4** The following table shows milligrams of seed planted, corresponding to the yield data in Table 7.6.

| | Variety | | | |
|---|---|---|---|---|
| Treatment | 1 | 2 | 3 | 4 |
| 1 | 2 | – | 7 | 3 |
| | 4 | | | 5 |
| | 3 | | | |
| 2 | 5 | 6 | – | – |
| | 3 | 4 | | |
| 3 | – | 6 | 6 | 4 |
| | | 2 | 8 | 6 |
| | | | | 5 |
| | | | | 7 |

Using this data and the data of Table 7.6 for covariance models (58)–(62), find the $b$'s and give the ANOVA Tables 8.3a and 8.3b.

**5** Consider equations (64) with the constraints $\alpha_1^\circ + \alpha_2^\circ = 0$ and $\beta_1^\circ + \beta_2^\circ = 0$. Show that the solution for $u^\circ$ is

$$u^\circ = \frac{2y_{1..} + 3y_{.3.} - y_{...}}{32}.$$

This would be the form of equation (63).

**6** Calculate the exact unbalanced data analyses for the data of Table 8.15 and compare them with Tables 8.17 and 8.19.

**7** The following data from Wright and Wilson (1979) gives the silt and clay content for three contiguous sites in Murcia, Spain. Two values are missing.

| Site | Silt or Clay | | | | | |
|---|---|---|---|---|---|---|
| 1 | Silt | 46.2 | 36.0 | 47.3 | $x$ | 30.9 |
| | Clay | 30.3 | 27.6 | 40.9 | 32.2 | 33.7 |
| 2 | Silt | 40.0 | 48.9 | 48.7 | 44.5 | 30.3 |
| | Clay | $y$ | 32.8 | 36.5 | 37.7 | 34.3 |
| 3 | Silt | 41.9 | 40.7 | 44.0 | 32.3 | 37.0 |
| | Clay | 34.0 | 36.6 | 40.0 | 30.1 | 38.6 |

Think of the two substances as nested within sites. Estimate the missing values and perform the ANOVA.

**8** Four different formulations of industrial glue are being tested. The tensile strength of the glue when it is applied to join parts is also related to the application thickness. Five observations on strength in pounds $y$ and thickness $z$ in 0.01 inches are obtained for each formulation. The data shown in the table below are taken from Montgomery (2005), and reproduced with kind permission from John Wiley & Sons.

| **Glue Formulation** | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 3 | | 4 | |
| $y$ | $z$ | $y$ | $z$ | $y$ | $z$ | $y$ | $z$ |
| 46.5 | 13 | 48.7 | 12 | 46.3 | 15 | 44.7 | 16 |
| 45.9 | 14 | 49.0 | 10 | 47.1 | 14 | 43.0 | 15 |
| 49.8 | 12 | 50.1 | 11 | 48.9 | 11 | 51.0 | 10 |
| 46.1 | 12 | 48.5 | 12 | 48.2 | 11 | 48.1 | 12 |
| 44.3 | 14 | 45.2 | 14 | 50.3 | 10 | 48.6 | 11 |

**(a)** Do the analysis of covariance and determine if at the 10% level of significance there is a significant difference in the tensile strength of the glue formulations. Estimate the *pooled within-class regression estimator.*

**(b)** Estimate the adjusted treatment means $\hat{m}_{iadj} = \bar{y}_{i.} - \hat{b}(\bar{z}_{i.} - \bar{z}_{..})$ for the tensile strengths for each of the four formulations.

**(c)** The standard error of the difference between two adjusted treatment means is given by

$$se = \left[ \text{MSE} \left( \frac{2}{n} + \frac{(\bar{y}_{i.} - \bar{y}_{j.})^2}{\text{SSE}_{zz}} \right) \right]^{1/2}$$

Find 90% Bonferonni simultaneous confidence intervals on the difference between formulations 1 and 2 and 3 and 4.

**9** The data below is due to Cameron and Pauling (1976). It is taken from Anderson and Herzberg (1985) and reproduced with kind permission of Springer Verlag. It compares the survival time, in days, of individual cancer patients who received vitamin C with the mean survival time of controls in days consisting of 10 patients who did not receive vitamin C. The survival time is that from first hospital attendance after the cancer reached the terminal stage. Three kinds of cancer are considered. The age of the patient in years is the covariate.

| Stomach | | | Bronchus | | | Colon | | |
|---|---|---|---|---|---|---|---|---|
| Age | Vitamin C | Control | Age | Vitamin C | Control | Age | Vitamin C | Control |
| 76 | 248 | 292 | 61 | 124 | 264 | 74 | 81 | 72 |
| 58 | 377 | 492 | 69 | 42 | 62 | 74 | 461 | 134 |
| 49 | 189 | 462 | 62 | 25 | 149 | 66 | 20 | 84 |
| 69 | 1843 | 235 | 66 | 45 | 18 | 52 | 450 | 98 |
| 70 | 180 | 294 | 63 | 412 | 180 | 48 | 246 | 48 |
| 68 | 537 | 144 | 79 | 51 | 142 | 64 | 166 | 142 |
| 50 | 519 | 643 | 76 | 1112 | 35 | 70 | 63 | 113 |
| 74 | 455 | 301 | 54 | 46 | 299 | 77 | 64 | 90 |
| 66 | 406 | 148 | 62 | 103 | 85 | 71 | 155 | 30 |
| 76 | 365 | 641 | 69 | 876 | 69 | 70 | 859 | 56 |
| 56 | 942 | 272 | 46 | 146 | 361 | 39 | 151 | 260 |
| 65 | 776 | 198 | 57 | 340 | 269 | 70 | 166 | 116 |
| 74 | 372 | 37 | 59 | 396 | 130 | 70 | 37 | 87 |
| 58 | 163 | 199 | | | | 55 | 223 | 69 |
| 60 | 101 | 154 | | | | 74 | 138 | 100 |
| 77 | 20 | 649 | | | | 69 | 72 | 315 |
| 38 | 383 | 162 | | | | 73 | 245 | 188 |

**(a)** Perform an analysis of covariance to determine whether

**1.** There is a significant difference in the survival time for patients who received vitamin C amongst the three types of cancer.

**2.** There is a significant difference in the survival time for patients in the control for the three types of cancer.

In each case, does the age of the patient affect the result?

**(b)** For each kind of cancer perform a paired $t$-test to determine whether patients who receive the vitamin C survive longer than the patients who do not. Also, do a $t$-test when the results for all three kinds of cancer are combined.

**10** The data below are concerned with the initial weights and growth rates of 30 pigs classified according to pen, sex, and type of food given. It is taken from Rao (1973, p. 291) and reproduced with the kind permission of John Wiley & Sons.

| Pen | Treatment | Sex | Initial Weight ($w$) | Growth Rate in Pounds per Week ($g$) |
|---|---|---|---|---|
| I | *A* | *G* | 48 | 9.94 |
| | *B* | *G* | 48 | 10.00 |
| | *C* | *G* | 48 | 9.75 |
| | *C* | *H* | 48 | 9.11 |
| | *B* | *H* | 39 | 8.51 |
| | *A* | *H* | 38 | 9.52 |
| II | *B* | *G* | 32 | 9.24 |
| | *C* | *G* | 28 | 8.66 |
| | *A* | *G* | 32 | 9.48 |
| | *C* | *H* | 37 | 8.50 |
| | *A* | *H* | 35 | 8.21 |
| | *B* | *H* | 38 | 9.95 |
| III | *C* | *G* | 33 | 7.63 |
| | *A* | *G* | 35 | 9.32 |
| | *B* | *G* | 41 | 9.34 |
| | *B* | *H* | 46 | 8.43 |
| | *C* | *H* | 42 | 8.90 |
| | *A* | *H* | 41 | 9.32 |
| IV | *C* | *G* | 50 | 10.37 |
| | *A* | *H* | 48 | 10.56 |
| | *B* | *G* | 46 | 9.68 |
| | *A* | *G* | 46 | 10.98 |
| | *B* | *H* | 40 | 8.86 |
| | *C* | *H* | 42 | 9.51 |
| V | *B* | *G* | 37 | 9.67 |
| | *A* | *G* | 32 | 8.82 |
| | *C* | *G* | 30 | 8.57 |
| | *B* | *H* | 40 | 9.20 |
| | *C* | *H* | 40 | 8.76 |
| | *A* | *H* | 43 | 10.42 |

**(a)** Perform the analysis of variance including the interaction between treatment and sex.

**(b)** Perform the analysis of covariance where the covariate (the initial weight of the pigs) is fitted first.

**(c)** What is the difference in terms of significance of the factor with and without considering the initial weight of the pigs?

**11** **(a)** Derive the distributions of $R(\mathbf{a})$ and SSRB shown in Table 8.3b.

**(b)** Show that $R(\mathbf{a})$, SSRB and $\mathbf{y}'\mathbf{y} - R(\mathbf{a}) -$ SSRB are pairwise independent.

**12** **(a)** For the general covariance model of Section 2a(i) and $\mathbf{P}$ of equation (10), prove that $\mathbf{Z}'\mathbf{P}\mathbf{Z}$ is non-singular.

**(b)** Using the result of (a), prove that $\mathbf{b}^\circ$ is estimable with respect to the model (6).

**(c)** Show that $\lambda'\mathbf{a}$ is estimable under the same conditions that it is estimable for the model without covariates.

**13** Show that

**(a)** The estimator $\mathbf{b}^\circ = \hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}\mathbf{y}$ is unbiased for $\mathbf{b}$.

**(b)** For estimable parametric functions, the estimator $\mathbf{p}'\mathbf{a}^\circ = \mathbf{p}'((\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{b}^\circ)$ is unbiased for $\mathbf{p}'\mathbf{a}$.

**(c)** If $(\mathbf{X}'\mathbf{X})^-$ is reflexive, the variance covariance matrix of $\mathbf{a}^\circ$ and $\mathbf{b}^\circ$ is

$$\begin{bmatrix} \text{var}(\mathbf{a}^\circ) & \text{cov}(\mathbf{a}^\circ, \mathbf{b}^\circ) \\ \text{cov}(\mathbf{a}^\circ, \mathbf{b}^\circ)' & \text{var}(\mathbf{b}^\circ) \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^- \sigma^2$$

What can you do if $(\mathbf{X}'\mathbf{X})^-$ is not reflexive?

**14** Graybill (1961, p. 392) gives the $F$-statistic for testing $H$: all $\alpha$'s equal in the one-way classification, with one covariate (in our notation)

$$\frac{1}{(c-1)\hat{\sigma}^2}\left\{\text{SSR}_{m,yy} + \text{SSE}_{yy} - \frac{(\text{SSR}_{m,yz} + \text{SSE}_{yz})^2}{\text{SSR}_{m,zz} + \text{SSE}_{zz}} - \left[\text{SSE}_{yy} - \frac{(\text{SSE}_{yz})^2}{\text{SSE}_{zz}}\right]\right\}.$$

Show the equivalence of this to $R(\alpha, \mu|b)$ of Table 8.4b.

**15** Derive an expression for SSE of Tables 8.3a and 8.3b which suggests that it is the residual sum of squares for fitting a linear model to $\mathbf{P}\mathbf{y}$. Describe the model.

**16** Show that the error sum of squares in Tables 8.4a and 8.4b is the same as that of fitting the model $\mathbf{y} - \hat{\mathbf{b}}\mathbf{z} = \mathbf{X}\mathbf{a} + \mathbf{e}$ for $\hat{\mathbf{b}}$ of (33) where the solution for $\mathbf{a}^\circ$ is that given before equation (32).

**17** Show that in Table 8.7, the statistic for testing the hypothesis $H$: $\alpha_i + b_i\bar{z}_{i.}$ equal for all $i$ is $R(\alpha|\mu)/(c-1)\hat{\sigma}^2$. [*Hint:* Use the result of the discussion in Sub-section 2a(vi) and Exercise 19 of Chapter 7.]

# 9

# INTRODUCTION TO VARIANCE COMPONENTS

The main interest in the models of Chapters 5–8 is estimation and tests of hypothesis about linear functions of the effects in the models. These effects are what we call fixed effects. The models are called fixed-effect models. However, there are situations where we have no interest in fixed effects. Instead, because of the nature of the data and the effects we are studying, our main interest would be in the variance. These kinds of effects are called random effects. The models involving them are called random-effects models. Models that involve a mixture of fixed effects and random effects are called mixed models.

The first major topic of this chapter is how to distinguish between fixed effects and random effects. We take this up by giving examples that illustrate the differences between these two kinds of effects. We emphasize the meaning and the use of these models in different situations without giving the mathematical details.

The variances associated with random effects are called variance components. The second major topic of this chapter is the estimation of variance components from balanced data. Chapter 10 deals with the more difficult topic of estimating variance components from unbalanced data.

## 1. FIXED AND RANDOM MODELS

Although the models of Chapters 5–8 are fixed-effects models, this is the first time we have referred to them as such. Therefore, our discussion of fixed and random effects begins with a fixed-effects model to confirm the use of this name.

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

### a. A Fixed-Effects Model

A classic experiment in agricultural research concerns testing the efficacy of nitrogen (N), potash (P), and potassium (K) on crop yield. Suppose an experiment of this kind involves 24 plants, with six plants receiving nitrogen, six plants getting potash, six plants potassium, and six plants getting no fertilizer at all, these being considered as control (C). A suitable model for analyzing the results of this experiment would be the one-way classification model (see Section 2 of Chapter 6). The model would then be

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{1}$$

where $y_{ij}$ is the $j$th observation on the $i$th treatment, with $\mu$ being a mean, $\alpha_i$ being the effect of the treatment $i$ and $e_{ij}$ an error term in the usual way.

Analysis of this experiment can lead to estimating $\alpha_1 - \alpha_4$, for example, and to testing the hypothesis $H\colon \alpha_1 - \alpha_4 = 0$. When studying differences of this nature, consider the treatments that we are dealing with. They are four very specific treatments of interest. In using them, we have no thought for any other kinds of fertilizer. Our sole interest is the study of N, P, and K in relation to each other and to no fertilizer. This is the concept of fixed effects. We fix our attention upon just the treatments of the experiment, upon these and no others. Thus, the effects are called *fixed effects*. Furthermore, because all the effects in the model are fixed effects (apart from the error terms which are always random), the model is called the *fixed-effects model*. It is often referred to as *Model I*, so named by Eisenhart (1947).

The inferences that we draw from data always depend on how we obtain the data. Therefore, we consider a sampling process that is pertinent to this fixed-effects model where the $\alpha$'s are the fixed effects of the four treatments, N, P, K, and C. We think of the data as one possible set of the data that we derive by repetitions of the experiment. On each occasion in these repetitions, the $e$'s are a random sample from a population of error terms distributed as $(\mathbf{0}, \sigma_e^2\mathbf{I})$. From this point on, we shall use $\sigma_e^2$ in place of $\sigma^2$ for the residual error variance. The randomness associated with the obtaining the $e$'s is what provides the means for making inferences about the functions of the $\alpha_i$'s and about $\sigma_e^2$.

### b. A Random-Effects Model

Suppose a laboratory experiment is designed to study the maternal ability of mice uses litter weights of 10-day-old litters as a measure of maternal ability, after the manner of Young et al. (1965). Six litters from each of the four dams (female parents), all of one breed, constitute the data. A suitable model for analyzing the data is the one-way classification model

$$y_{ij} = \mu + \delta_i + e_{ij}. \tag{2}$$

For this model, $y_{ij}$ is the weight of the $j$th litter from the $i$th dam. The effect due to the $i$th dam is $\delta_i$ and $e_{ij}$ is the customary error term.

Consider the $\delta_i$'s and the dams they represent. The data relate to maternal ability. This is a variable that is certainly subject to biological variation from animal to animal. Therefore, the prime concern of the experiment will probably not center on specifically the four female mice used in the experiment. After all, they are only a sample from a large population of mice, the females of the breed. Each of these females has some ability in a maternal capacity. Therefore, the animals that are in the experiment are thought of as a random sample of four from a population of females.

In the fertilizer experiment previously described, each fertilizer is of specific importance and interest, with no thought of it being a sample from a population of fertilizers. However, in the mouse experiment, each mouse is merely a sample (of one) from a population of female mice. Nothing important has conditioned our choice of one mouse over the other. We have no specific interest in the difference between any one of our four mice and any other of them. However, interest does lie in the extent to which maternal ability varies throughout the population of mice. It is to this end that our model is directed.

The sampling process involved in obtaining the mouse data is taken as being such that any one of the many possible sets of data could be derived from repetitions of the data gathering process. By concentrating attention on repetitions, we do not confine ourselves to always having the same four mice. We imagine getting a random sample of four on each occasion from the population of mice. In addition, whatever four mice we get on any occasion, we think about getting a random sample of $e$'s from a population of errors as was the case with the fixed-effects model. Our concept of error terms is the same for both the fixed-effects model and the random-effects model. The important difference between the two models is that in the fixed-effects model, we conceive of always having the same $\alpha$'s, the same treatments, while in the random-effects model, the mice data, we think of taking a random sample of mice on each occasion. Thus, the $\delta_i$'s of our data are a random sample from a population of $\delta$'s. Insofar as the data are concerned, the $\delta_i$'s therein are random variables. In this context, we call them *random effects*. Correspondingly, the model is called the *random-effects model* or, sometimes, the *random model*. Eisenhart (1947) called it *Model II*. This name continues to receive widespread use.

In each model, the error terms are a random sample from a population distributed as $(\mathbf{0}, \sigma_e^2\mathbf{I})$. However, for the fixed-effects model, the $\alpha$'s represent the effects of specific treatments while in the random model the $\delta$'s are also a random sample from a population that is distributed as $(\mathbf{0}, \sigma_\delta^2\mathbf{I})$. In addition, sampling of the $\delta$'s is assumed to be independent of that of the $e$'s and so covariances between the $\delta$'s and the $e$'s are zero. Furthermore, if the distribution of the $\delta$'s was to have a non-zero mean $\mu_\delta$, we could rewrite the model (2) as

$$y_{ij} = (\mu + \mu_\delta) + (\delta_i - \mu_\delta) + e_{ij}. \tag{3}$$

Thus, if we define $\mu + \mu_\delta$ as the mean and $\delta_i - \mu_\delta$ as the dam effect, the latter would have zero mean. Therefore, there is no loss of generality in taking the mean of the $\delta$'s to be zero.

With the $\delta$'s and the $e$'s of (2) being random variables with variances $\sigma_\delta^2$ and $\sigma_e^2$, respectively, from (2), the variance of an observation is $\sigma_y^2 = \sigma_\delta^2 + \sigma_e^2$. Accordingly, the variances $\sigma_\delta^2$ and $\sigma_e^2$ are called *variance components*. The model is sometimes referred to as a *variance component model*. The objectives of using such a model are the estimation of the variance components and inferences about them.

### c. Other Examples

**(*i*) *Of Treatments and Varieties.*** The fixed-effects model of equation (1) relates to four fertilizer treatments. Suppose we expand this experiment to using each of four treatments on six different plants of each of three varieties of the plant. A suitable model would be the two-way classification model with interaction. The observations $y_{ijk}$ represent the yield of the $k$th plant of the $j$th variety receiving the $i$th treatment. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \tag{4}$$

For this model, $\mu$ is a general mean, $\alpha_i$ is the effect of the $i$th treatment on the yield, $\beta_j$ is the effect of the $j$th variety, $\gamma_{ij}$ is the interaction, and $e_{ijk}$ is the usual error term. Just as the treatment effects $\alpha_i$ were earlier described as fixed effects, so they are now. Likewise, the variety effects $\beta_j$ are also fixed effects. In this experiment, interest in varieties centers solely on the three varieties that we use. There is no thought that they are a random sample from some population of varieties. Thus, we consider both the $\alpha_i$ and $\beta_j$ and their interactions as fixed effects and we have a fixed-effects model.

**(*ii*) *Of Mice and Men.*** Suppose the mouse experiment had been supervised by three laboratory technicians, one for each successive pair of litters that the mice had. One possible model for the resulting data would be

$$y_{ijk} = \mu + \delta_i + \tau_j + \theta_{ij} + e_{ijk}. \tag{5}$$

The observation $y_{ijk}$ represents the weight of the $k$th litter from the $i$th dam being cared for by the $j$th technician. The effect on litter weight of the $i$th dam is $\delta_i$. The interaction is $\theta_{ij}$. Earlier, we explained how $\delta_i$ is a random effect, representing the maternal capacity of the $i$th dam chosen randomly from a population of (female) mice. It is not difficult to imagine $\tau_j$ as being a random effect of similar nature. A laboratory experiment has to be cared for. Usually, there is little interest as far as the experiment itself is concerned in who the technician tending to it is. Reasonably, one can think of him/her as a random sample of one from some population of laboratory technicians. Thus, in the whole experiment, we have a random sample of three technicians. Correspondingly, the $\tau_j$ are random effects with zero mean and variance $\sigma_\tau^2$. Likewise, the interaction effects are also random, with zero mean and variance $\sigma_\theta^2$. All covariances are taken as zero. Thus, all elements of the model (5)—save $\mu$—are random effects and we have a random model. Apart from $\mu$, the parameters

of interest are $\sigma_\delta^2$, $\sigma_\tau^2$, and $\sigma_\theta^2$. These represent the influence of dam, technician, and dam by technician, respectively, on the variance of $y$. The part of the variance that is not accounted for by these effects is $\sigma_e^2$, the residual variance, in the usual manner.

**(*iii*) *Of Cows and Bulls*** Another example of the random model arises in dairy cow breeding. With the advent of artificial insemination, a bull can sire offspring in many different places simultaneously and have progeny in numerous different herds. When the females among these progeny themselves calve and start to give milk, analyses of their milk yields can be made. A suitable model for $y_{ijk}$, the milk yield of the $k$th daughter in herd $i$ sired by bull $j$, is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \tag{6}$$

The effect of the cow's being in herd $i$ is $\alpha_i$, $\beta_j$ is the effect of bull $j$, $\gamma_{ij}$ is the interaction effect, and $e_{ijk}$ is the customary random error term. In this case, all of the effects are considered random. The herds involved in the data are assumed to be a random sample from a population of herds. The bulls are taken to be from a random sample of bulls. The interaction effects are also assumed to be random. These effects are assumed to be mutually independent, with variances $\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$, and $\sigma_e^2$, respectively. The animal breeder is interested in estimating these variances so that he/she can estimate the ratio $4\sigma_\beta^2/(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2)$. This ratio is important in bringing about increased milk production through selective breeding.

## 2. MIXED MODELS

A general mean $\mu$ (a fixed effect) and error terms $e$ (random) occur in all the preceding examples, as they do in most models. Apart from these, all effects in each of the preceding models are either fixed or random. We now consider models where some of the effects (other than $\mu$ and $e$) are fixed and some are random. Such models are called *mixed models*. Of course, any model containing a fixed effect $\mu$ and random error terms is truly a mixed model. However, the description of mixed models is usually reserved for situations where effects other than $\mu$ and $e$ are a mixture of fixed and random effects. In some situations as we shall see (Section 8 of Chapter 10), it is convenient to treat all models as if they are mixed models. Generally, however, the distinction is made between fixed, random, and mixed models as described here. We now give some examples of mixed models.

**(*i*) *Of Mice and Diets.*** Suppose in the mouse experiment that instead of the mice being cared for by three different technicians, one man supervised the whole experiment. Suppose, further, that three specially prepared diets were used, with the purpose

of the experiment being to compare the three diets. Then, if $y_{ijk}$ is the $k$th litter weight of the $i$th dam when receiving diet $j$,

$$y_{ijk} = \mu + \delta_i + \varphi_j + \gamma_{ij} + e_{ijk}. \tag{7}$$

Now, though, because the diets are three specific diets of interest, the $\varphi_j$ effects representing those diets are fixed effects. As before, the $\delta_j$—the dam effects—are random. Thus, (7) is a model containing fixed effects $\varphi_j$ and random effects $\delta_j$. This is a *mixed model*, a mixture of fixed and random effects.

Notice that (7) contains interaction effects $\gamma_{ij}$ for interactions between dams and diets. Since dams are being taken as random effects, it is logical that these interactions are random also. Thus, the model has $\varphi_j$ as fixed effects and the $\delta_j$ and $\gamma_{ij}$ as random, having zero means and variances $\sigma_\delta^2$ and $\sigma_\gamma^2$, respectively.

**(*ii*) *Of Treatments and Crosses*** In an experiment concerning fertilizers, suppose that six plants of each of 20 replicate crosses of two varieties of the crop (early and late-ripening tomatoes, say) are used. Each cross would be a random sample from the infinite number of times the two varieties could be crossed. The equation for the model could still be equation (4). However, $\beta_j$ would now be a random effect for the $j$th replicated cross. The $\gamma_{ij}$ would be the (random) interaction effect between the $i$th fertilizer treatment and the $j$th cross. Thus, equation (4), formerly appropriate to a fixed-effects model, is now suited to a mixed model. The equation of the model is unchanged but the meanings of some of its terms have changed.

**(*iii*) *On Measuring Shell Velocities*** Thompson (1963), following Grubbs (1948), discusses the problem of using several instruments to measure the muzzle velocity of firing a random sample of shells from a manufacturer's stock. A suitable model for $y_{ij}$ the velocity of the $i$th shell measured by the $j$th measuring instrument, is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}.$$

In this model, $\alpha_i$ is the effect of the $i$th shell and $\beta_j$ is the bias in instrument $j$. Since the shells fired are a random sample of shells the $\alpha_i$'s are random effects. The $\beta_j$ are fixed effects because the instruments used are the only instruments of interest.

**(*iv*) *Of Hospitals and Patients*** The following experiment was discussed by Igor Ruczinski, an Associate Professor in The Department of Biostatistics at The Johns Hopkins University School of Public Health. The results are the basis for Exercise 19.2 of Gruber (2014, p. 263). Suppose we have three specific hospitals, four randomly chosen patients within each hospital and two independent measurements of blood coagulation in seconds. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}, \quad i = 1, 2, 3, \quad j = 1, 2, 3, 4, \quad \text{and} \quad k = 1, 2..$$

The $y_{ijk}$ represents the $k$th measurement of blood coagulation time for the $j$th patient in the $i$th hospital. The $\alpha_i$ represents the effect of the $i$th hospital. The $\beta_{ij}$ represents the effect of the $j$th patient within the $i$th hospital. The factor patient is nested within the hospitals, so this is a nested model. Since we have three specific hospitals, $\alpha_i$ is a fixed effect. Since within each hospital the patients are chosen at random, $\beta_{ij}$ is a random effect and again we have a mixed model. The variance components are $\sigma_\alpha^2, \sigma_{\beta(\alpha)}^2$, and $\sigma_e^2$. The subscript $\beta(\alpha)$ on the second term indicates that the patients are nested within the hospitals.

## 3. FIXED OR RANDOM

Equation (4) for the treatments and varieties example is indistinguishable from (6) for the bull and herds example. However, the models involved in the two cases are different because of the interpretation attributed to the effects. In the treatments and varieties example, they are fixed. In the bulls and herds example, they are random. In these and other examples that we discuss, most of the effects are categorically fixed or random. Fertilizer treatments are fixed effects, as are diets and measuring instruments. Likewise, mice, bulls, and artillery shells are random effects. How about the laboratory technicians, where three of them cared for the mice; or the herds where the bull's progeny were being milked? In each case, we assumed that these effects were random. However, this might not always be the case. For example, each one of the technicians may not have come from a random sample of employees. Maybe all of them were available and we wanted to assess differences between three specific technicians. In that case, the technician effects in equation (5) would be fixed effects, not random. The same might be true about the herd effects in equation (6). Typically, analyses of such data usually involve hundreds of sales that are considered a random sample from a larger population. However, if the situation was one of analyzing just a few herds, five or six, say wherein the sole interest lay in just those herds, then the herd effects in (6) would be more appropriately fixed and not random. Thus, the deciding factor that determines whether the effects of a factor are fixed or random is the situation to which the model applies.

There are situations where deciding whether certain effects are fixed or random is not immediately obvious. For example, consider the case of year effects in studying wheat yields. Are the effects of years on yields to be considered fixed or random? The years themselves are unlikely to be random because they probably will be a group of consecutive years over which the data may have been gathered or the experiments run. However, the effects on the yield may reasonably be considered random—unless, perhaps, one is interested in comparing specific years for some purpose.

When attempting to decide whether effects are fixed or random in the context of the data, the determining factors are the manner in which the data were gathered and the environment it came from. The important question is for what levels of the factors under consideration are inferences to be drawn? If the inferences were to be just for the specific factors being thought about, then the effects would be considered fixed. On the other hand, if the inferences are being made not just about the levels occurring

in the data but about some population from which the data are considered to be a random sample, then the effects would be considered as random.

We emphasize that the assumption of randomness does not include with it the assumption of normality. This assumption is frequently made for random effects. However, it is a separate assumption, made after assuming that the effects are random. Although most estimation procedures for variance components do not require normality, the normality of random effects is often assumed when distributional properties of the estimators are investigated.

## 4. FINITE POPULATIONS

We assume that random effects occurring in data are from a population of effects. Usually, we assume that the populations have infinite size like, for example, the population of all possible crosses between two varieties of tomato. They could be crossed an infinite number of times. However, the definition of random effects does not demand infinite populations of such effects. They can be finite. Furthermore, finite populations can be very large, so large that they can be considered infinite for all practical purposes. For example, consider the population of all mice in New York State on July 4, 2015! Hence, random effects factors can have conceptual populations of three kinds insofar as their size is concerned: infinite, finite but so large as to be deemed infinite and finite.

We shall concern ourselves with random effects coming solely from populations that we assume are infinite either because this is the case or because, although finite, the population is large enough to be taken as infinite. These are the situations that occur most often in practical problems. Discussion of finite populations, in particular, variance components, may be found in several places. See for example, Bennett and Franklin (1954, p. 404), Gaylor and Hartwell (1969), and Sahai (1974). Searle and Fawcett (1970) give rules for converting the estimation procedure of any infinite-population situation into one of finite populations.

## 5. INTRODUCTION TO ESTIMATION

We consider an important and frequently used method for the estimation of variance components in balanced data. For unbalanced data, there are a number of methods available that simplify to the method that we are about to present for balanced data. For this reason, we consider balanced data first. The method of estimating variance components for any random or mixed model relies on the mean squares of the analysis of variance for the corresponding fixed-effects model. The general procedure consists of calculation of the analysis of variance as if the model were a fixed-effects model and then deriving the expected values of the mean squares under the random or mixed model. Certain of the expected values will be linear functions of the variance components. Equating these expected mean squares to their calculated (observed) values leads to linear equations in the variance components. The solutions to these

linear equations are taken to be the estimators of the variance components. This method of estimating variance components is known as the *analysis of variance method*.

Mean squares in analysis of variance are quadratic forms in the observations. Therefore, we can derive their expected values from Theorem 4 of Chapter 2, wherein $\mathbf{V}$ is the variance–covariance matrix of the observations. Although for balanced data, this is not the easiest method for calculating the expected values for mean squares, it is instructive to demonstrate the form of the $\mathbf{V}$-matrix for a simple random model. It is the basis of such matrices for unbalanced data for which Theorem 4 of Chapter 2 is of utmost importance. We illustrate this by means of the mouse example of Section 1b.

### a. Variance Matrix Structures

In all the fixed-effects models of Chapters 5–8, the covariance matrix of the observations var($\mathbf{y}$) has been of the form $\sigma_e^2\mathbf{I}_N$. However, the form of the covariance matrix for random and mixed models is different because the covariance structure of the random effects is what determines the variance–covariance matrix of the vector of observations.

Suppose we rewrite the model for the mouse example, equation (2) as

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{8}$$

where $\mu$ and $e_{ij}$ are the same as in (2) and $\alpha_i$ is now used in place of $\delta_i$. Thus, $\alpha_i$ is a random effect with zero mean and variance $\sigma_\alpha^2$. It is independent of the $e$'s and the other $\alpha$'s. Thus, we have that $E(\alpha_i\alpha_k) = 0$ for $i \neq k$ and $E(\alpha_i e_{i'j'}) = 0$ for all $i$, $i'$, and $j'$. From this we have,

$$\operatorname{cov}(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_\alpha^2 + \sigma_e^2 & \text{for} \quad i = i' \quad j = j' \\ \sigma_\alpha^2 & \text{for} \quad i = i' \quad j \neq j' \\ 0 & \text{for} \quad i \neq i'. \end{cases}$$

Hence, for example, the variance–covariance matrix for the matrix of six observations on the first dam is

$$\operatorname{var}\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \end{bmatrix} = \begin{bmatrix} \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_e^2 \end{bmatrix}$$

$$= \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J}. \tag{9}$$

We meet this form of matrix repeatedly: $\lambda_1\mathbf{I} + \lambda_2\mathbf{J}$, where $\lambda_1$ and $\lambda_2$ are scalars, usually variances, and $\mathbf{J}$ is a square matrix with every element unity. In the present case, it is the covariance matrix of the set of six litter weights from each dam. Since the weights are independent, as between one dam and another, the covariance matrix of all 24 weights can be partitioned as

$$\text{var}(\mathbf{y}) = \begin{bmatrix} \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J} \end{bmatrix},$$

where $\mathbf{I}$ and $\mathbf{J}$ have order equal to the number of observations in the classes, in this case 6. Using $\Sigma^+$ to denote the operation of direct sum, as in Section 2a of Chapter 6, we write

$$\text{var}(\mathbf{y}) = \sum_{i=1}^{4}{}^{+}(\sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J}). \tag{10}$$

We will make frequent use of the notation in (10), especially with unbalanced data, in the form $\sum_{i=1}^{a}{}^{+}(\sigma_e^2\mathbf{I}_i + \sigma_\alpha^2\mathbf{J}_i)$ where $\mathbf{I}_i$ and $\mathbf{J}_i$ have order $n_i$.

### b. Analyses of Variance

The one-way classification model of Section 2d of Chapter 6 is suitable for the fertilizer experiment discussed in Section 1a. We show its analysis of variance in Table 9.1 based on Table 6.4.

The basic use of Table 9.1 is to summarize calculation of the $F$-statistic $\text{MS}_\text{m}/\text{MSE}$ for testing $H$: all $\alpha$'s equal. The lower section of the table contains the expected value of the mean squares. We usually do not show this for the fixed-effects model. Nevertheless, its presence emphasizes the hypothesis that can be tested by the $F$-statistic.

This is true because, for $F = Q/s\hat{\sigma}^2$ used so much in the earlier chapters,

$$F \sim F'\left[s, N-r, \frac{[E(Q) - s\sigma^2]}{2\sigma^2}\right].$$

We can show this by applying Theorems 4 and 5 of Chapter 2 to $Q$. Therefore, the hypothesis concerning $s$ LIN estimable functions that makes $[E(Q) - s\sigma^2]$ zero is tested by comparing $F = Q/s\hat{\sigma}^2$ against the central $F(s, N-r)$ distribution.

**TABLE 9.1 Analysis of Variance for Four Fertilizer Treatments Each Used on Six Plants**

| Source of Variation | d.f. | Sum of Squares | Mean Squares |
|---|---|---|---|
| Mean | 1 | $\text{SSM} = R(\mu) = 24\bar{y}_{\cdot\cdot}^2$ | MSM = SSM/1 |
| Treatments | 3 | $\text{SSR}_\text{m} = R(\alpha\|\mu) = \sum_{i=1}^{4} 6(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ | $\text{MSR}_\text{m} = \text{SSR}_\text{m}$ / 3 |
| Residual error | 20 | $\text{SSE} = \text{SST} - R(\mu, \alpha) = \sum_{i=1}^{4}\sum_{j=1}^{6} (y_{ij} - \bar{y}_{i\cdot})^2$ | MSE = SSE/20 |
| Total | 24 | $\text{SST} = \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2$ | |

| Expected mean squares |
|---|
| $E(\text{MSM}) = 24\left(\mu + \frac{1}{4}\sum_{i=1}^{4}\alpha_i\right)^2$ |
| $E(\text{MSR}_\text{m}) = \frac{6}{3}\sum_{i=1}^{4}\left(\alpha_i - \frac{1}{4}\sum_{i=1}^{4}\alpha_i\right)^2 + \sigma_e^2$ |
| $E(\text{MSE}) = \sigma_e^2$ |

**Example 1 Test of Hypothesis in Table 9.1** In Table 9.1

$$E(\text{SSR}_\text{m}) = 6\sum_{i=1}^{4}(\alpha_i - \bar{\alpha}_{\cdot})^2 + 3\sigma_e^2.$$

Hence $F = \text{SSR}_\text{m}/3\sigma_e^2$ tests the hypothesis that makes $6\sum_{i=1}^{4}(\alpha_i - \bar{\alpha}_{\cdot})^2$ zero, namely,

$$H\colon \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4.$$ □

Expected mean squares are useful in indicating the hypotheses tested by the corresponding $F$-statistic. However, the reason we show them in Table 9.1 is for comparison with the random model case of the mouse experiment in Section 1a. The fixed effects analogue of the model for the mouse experiment is the same as that of the fertilizer experiment. The variance components for the mouse experiment are estimated from the analysis of variance in Table 9.2. Except for the expected values on mean squares, Table 9.1 and 9.2 are identical. In both cases, we can obtain these expected values from Theorem 4 of Chapter 2. For Table 9.1, $\mathbf{V} = \text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$. For Table 9.2, $\mathbf{V}$ is given by (10). An alternative (and often easier) derivation is the "brute

**TABLE 9.2 Analysis of Variance of Four Dams Each Having Six Litters**

| Source of Variation | d.f. | Sum of Squares | Mean Squares |
|---|---|---|---|
| Mean | 1 | $\text{SSM} = R(\mu) = 24\bar{y}_{..}^2$ | MSM = SSM/1 |
| Treatments | 3 | $\text{SSR}_\text{m} = R(\alpha\|\mu) = \sum_{i=1}^{4} 6(\bar{y}_{i.} - \bar{y}_{..})^2$ | $\text{MSR}_\text{m} = \text{SSR}_\text{m}$ / 3 |
| Residual error | 20 | $\text{SSE} = \text{SST} - R(\mu, \alpha) = \sum_{i=1}^{4}\sum_{j=1}^{6} (y_{ij} - \bar{y}_{i.})^2$ | MSE = SSE/20 |
| Total | 24 | $\text{SST} = \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2$ | |

| Expected mean squares | |
|---|---|
| $E(\text{MSM}) =$ | $24\mu^2 + 6\sigma_\alpha^2 + \sigma_e^2$ |
| $E(\text{MSR}_\text{m}) =$ | $6\sigma_\alpha^2 + \sigma_e^2$ |
| $E(\text{MSE}) =$ | $\sigma_e^2$ |

force" one of substituting the equation of the model into the mean squares and then taking expectations using the appropriate model in each case. In practice, we do not have to use either of these methods for balanced data. Simple rules of thumb apply, as we shall see in Section 6. We do not illustrate either method here. We give an illustration for the two-way balanced data in Section 7 and for unbalanced data in Chapter 10.

### c. Estimation

The residual error variance in the fixed-effects model of Table 9.1 is estimated in the usual way by $\hat{\sigma}_e^2 = \text{MSE}$. This is tantamount to the analysis of variance method of estimating variance components by equating mean squares to their expected values. We continue it in Table 9.2 to obtain not only

$$\hat{\sigma}_e^2 = \text{MSE} \quad \text{but also} \quad 6\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2 = \text{MSR}_\text{m}.$$

The solutions to these equations are

$$\hat{\sigma}_e^2 = \text{MSE} \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{(\text{MSR}_\text{m} - \text{MSE})}{6}.$$

These are the estimators of $\sigma_e^2$ and $\sigma_\alpha^2$.

The preceding example is the simplest illustration of estimating variance components from balanced data of a random (or mixed) model. It extends easily to other

balanced data situations. In the analysis of variance, there will be as many mean squares whose expectations do not involve fixed effects, as there are variance components to be estimated. Equating each of these mean squares to their expected values gives a set of linear equations in the variance components. The solution to these linear equations is the estimators of the variance components. For example, in Table 9.2, $E$(MSM) involves $\mu$. The other expected mean squares do not and so they yield the estimators of the variance components of the model. For random models, the only expected mean square that involves fixed effects is $E$(MSM), that for means. In mixed models, there will also be others. However, there will also be sufficient expected mean squares that do not involve fixed effects to provide equations that yield estimators of the variance components. This is the analysis of variance method of estimating variance components.

**Example 2 An Example of Estimating Variance Components** The data for this example are taken from a large industrial experiment performed at Eastman Kodak Company, Rochester, New York. It was obtained by courtesy of Dr. James Halavin, Professor Emeritus, School of Mathematical Sciences, Rochester Institute of Technology. These data and example is also discussed in Gruber (2010).

Six different units are chosen at random from a large number of units of a certain type of camera. For each unit, the time from the first flash of the camera until the camera's ready light went back on was measured. Six readings were taken for each camera. The data are below.

| Camera | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | 4.39 | 4.16 | 4.89 | 5.00 | 6.53 | 5.71 |
| | 4.34 | 4.88 | 4.78 | 4.45 | 5.38 | 4.94 |
| | 4.61 | 6.81 | 5.16 | 5.00 | 6.21 | 5.71 |
| | 4.56 | 6.53 | 5.94 | 4.50 | 5.77 | 5.71 |
| | 5.89 | 5.22 | 5.44 | 5.54 | 6.97 | 6.26 |
| | 4.12 | 4.16 | 4.67 | 4.56 | 5.54 | 4.66 |

We wish to determine whether there is significant variability amongst camera units in the time from first flash until the ready light comes back on. Consider the SAS output below.

**The SAS System**

**The GLM Procedure**

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **camera** | 6 | 1 2 3 4 5 6 |
| **Number of Observations Read** | | 36 |
| **Number of Observations Used** | | 36 |

**The SAS System**

**The GLM Procedure**
**Dependent Variable: time**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 5 | 9.79948889 | 1.95989778 | 6.27 | 0.0004 |
| **Error** | 30 | 9.38300000 | 0.31276667 | | |
| **Corrected Total** | 35 | 19.18248889 | | | |
| **R-Square** | **Coeff Var** | | **Root MSE** | **time Mean** | |
| 0.510856 | 10.68414 | | 0.559255 | 5.234444 | |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **camera** | 5 | 9.79948889 | 1.95989778 | 6.27 | 0.0004 |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **camera** | 5 | 9.79948889 | 1.95989778 | 6.27 | 0.0004 |

**The SAS System**

**The GLM Procedure**

| **Source** | **Type III Expected Mean Square** |
|---|---|
| **camera** | Var(Error) + 6 Var(camera) |

The program to generate this output is similar to that used before with the additional command `random camera;` after the model statement.

We see from the computer output that at the 1% level of significance, there is indeed a significant variability, that is, we reject the hypothesis $H\colon \sigma_\alpha^2 = 0$ at $\alpha = .01$, the $p$-value being $0.0004 < .01$. To estimate the variance components, we have the equations

$$\hat{\sigma}_e^2 = .3128$$
$$6\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2 = 1.960.$$

Then,

$$\hat{\sigma}_\alpha^2 = \frac{1.960 - 0.313}{6} = 0.2745. \qquad \square$$

The procedure of "equating mean squares to their expected values" is a special case of the more general procedure of equating quadratic forms to their expected values, as used in a variety of ways with unbalanced data. These are discussed in Chapter 10. For balanced data, the "obvious" quadratic forms to use are the analysis of variance mean squares. It turns out that the resulting estimators have several optimal properties. Since derivation of the estimators depends on the availability of expected mean squares, we turn first to these and the rules that enable them to be written down on sight. Subsequently, we consider the properties of the estimators.

## 6. RULES FOR BALANCED DATA

We confine discussion to factorial designs, consisting of crossed and nested classifications and combinations thereof, where the number of observations in all of the sub-most subclasses is the same. We exclude situations of partially balanced data, such as in Latin squares, balanced incomplete blocks and their extensions. Otherwise, the rules of thumb for setting up analysis of variance tables apply to any combination or any number of crossed and/or nested classifications. These rules lay out procedures for determining:

(i) the lines in the analysis of variance;
(ii) their degrees of freedom;
(iii) formulae for calculating sums of squares;
(iv) expected values of mean squares.

Most of the rules are based on Henderson (1959, 1969). Rule 9 is an exception. It comes from Millman and Glass (1967). They rely heavily on the Henderson paper for a similar set of rules.

The description of the rules is purposefully brief with no attempt at substantiation. However, justification of the rules is available in Lum (1954) and Schultz (1955).

### a. Establishing Analysis of Variance Tables

**(*i*) *Factors and Levels.*** The analysis of variance table is described in terms of factors $A, B, C,\ldots$, with the number of levels in them being $n_a, n_b, n_c, \ldots$, respectively. When one factor is nested within another, the notation will be $C$: $B$ for factor $C$ within $B$, $C$: $BA$ for $C$ within $AB$ subclasses, and so on. A letter on the left of the colon represents the nested factor and those on the right of the colon represent the factors in which the nested factor is found. For example, for a nested factor $C$, $n_c$ is the number of levels of factor $C$ within each of the factors in which it is nested. Factors that are not nested, namely those forming cross-classifications will be called crossed factors.

Within every sub-most class of the data, we assume that there are the same number of observations $n_w$, either one or more than one. In either case, these observations can, as Millman and Glass (1967) point out, be referred to as replications within all other subclasses. Following Henderson (1959), we refer to these as the "within" factor using the notation $W$: $ABC\ldots$, the number of levels of the "within" (i.e., number of replicates) being $n_w$. The total number of the observations is then the product of the $n$'s; to be specific

$$N = n_a n_b n_c \ldots n_w.$$

**(*ii*) *Lines in the Analysis of Variance Table***

**Rule 1**. There is one line for each factor (crossed or nested), for each interaction and for "within."

(*iii*) ***Interactions.*** Interactions are obtained symbolically as the product of factors, both factorial and nested. We consider all products of 2, 3, 4,… factors. For the sake of generality, we assume that all crossed factors have a colon on the right of the symbol; for example, *A*:, *B*:, and so on.

**Rule 2**. Every interaction is of the form *ABC*…: *XYZ*…, where *ABC*… is the product on the left of the colon of the factors being combined and *XYZ*… is the product on the right of the colon of the factors so associated with *A*, *B*, and *C*…

**Rule 3**. Repeated letters on the right of the colon are replaced by one of their kind.

**Rule 4.** If any letter occurs on both sides of a colon, that interaction does not exist.

**Example 3 Illustrations of Rules 2–4**

| Factors | Interaction | |
|---|---|---|
| *A* and *B* | *AB* | (Rule 2) |
| *A* and *C*: *B* | *AC*: *B* | (Rule 2) |
| *A*: *B* and *C*: *B* | *AC*: *BB* = *AC*: *B* | (Rule 3) |
| *A*: *B* and *B*: *DE* | *AB*: *BDE* nonexistent | (Rule 4) |

The symbolic form *W*: *ABC*… for replicates does, by Rule 4, result in no interactions involving *W*. Furthermore, the line in the analysis of variance labeled *W*: *ABC*…, being the "within" line, is the residual error line. □

(*iv*) ***Degrees of Freedom*** Each line in an analysis of variance table refers either to a crossed factor (such as *A*:), to a nested factor (such as *C*: *B*) or to an interaction (e.g., *AC*: *B*). Therefore, any line may be typified by the general expression given for an interaction in Rule 2, namely *ABC*…: *XYZ*…

**Rule 5.** Degrees of freedom for the line are denoted by

$$AB\text{: }XY \text{ are } (n_a - 1)(n_b - 1)n_x n_y.$$

The rule is simple. Degrees of freedom are the product of terms like $(n_a - 1)$ for every letter *A* on the left of the colon and of terms like $n_x$ for every letter *X* on the right of the colon.

**Rule 6.** The sum of all degrees of freedom is $N - 1$, with $N = n_a n_b n_c \ldots$

(*v*) ***Sums of Squares*** We use the symbols that specify a line in the analysis of variance to establish the corresponding sum of squares. We take the basic elements to be the uncorrected sum of squares with notation:

$$1 = CF = N\bar{y}^2$$

and *a*, *ab*, *abc* ≡ uncorrected sums of squares for the *A* factor, the *AB*, and the *ABC* subclasses, respectively.

**Rule 7**. The sum of squares for the line denoted by

$$AB{:}\,XY \text{ is } (a-1)(b-1)xy = abxy - axy - bxy + xy.$$

The rule is again simple. Symbolically, a sum of squares is the product of terms like $(a - 1)$ for every letter $A$ on the left of the colon and of terms like $x$ for every letter $X$ on the right of the colon. This rule is identical to Rule 5 for degrees of freedom. If in the expression for degrees of freedom every $n_f$ is replaced by $f$, the resulting expression is, symbolically, the sum of squares. For example,

$$(n_a - 1)(n_b - 1)n_x n_y \text{ becomes } (a-1)(b-1)xy = abxy - axy - bxy + xy.$$

After expansion, interpretation of these products of lower case letters is as uncorrected sums of squares.

Observe that all sums of squares are expressed essentially in terms of crossed factors. Even when a factor is nested, sums of squares are expressed in terms of uncorrected sums of squares calculated as if the nested factor were a crossed factor. For example, the sum of squares for $A{:}B$ ($A$ within $B$) is $(a - 1)b = ab - b$, where $ab$ is the uncorrected sum of squares of the $AB$ subclasses.

**Rule 8**. The total of all sums of squares is $\sum y^2 - CF$ where $\sum y^2$ represents the sum of squares of the individual observations, $wabc\ldots$ in the above notation, and where $CF$ is the correction factor.

**Example 4 Illustrations of Rules 1–8** Table 9.3 shows the analysis of variance that we derive from these rules for the case of two crossed classifications $A$ and $B$, a classification $C$ nested within $B$, namely $C{:}\,B$ and the within factor $W{:}\,ABC$. Application of these rules is indicated at the appropriate points in the table.

**TABLE 9.3 Example of Rules 1–8: Analysis of Variance for Factors $A$, $B$, $C{:}\,B$, and $W{:}$ AB**

| Line (Rules 1–4) | Degrees of Freedom (Rule 5) | Sum of Squares (Rule 7) |
|---|---|---|
| $A$ | $n_a - 1$ | $(a-1) = a - 1$ |
| $B$ | $n_b - 1$ | $(b-1) = b - 1$ |
| $C{:}\,B$ | $(n_c - 1)n_b$ | $(c-1)b = bc - b$ |
| $AB$ | $(n_a - 1)(n_b - 1)$ | $(a-1)(b-1) = ab - a - b + 1$ |
| $AC{:}\,B$ | $(n_a - 1)(n_c - 1)n_b$ | $(a-1)(c-1)b = abc - ab - bc + b$ |
| $W{:}\,ABC$ | $(n_w - 1)n_a n_b n_c$ | $(w-1)abc = wabc - abc$ |
| Total | $N - 1$ (Rule 6) | $\sum y^2 - CF = wabc$ (Rule 8) |

□

### b. Calculating Sums of Squares

So far the uncorrected sums of squares denoted by lower case letters such as $a$ and $ab$ have been defined solely in words. For example, $ab$ is the uncorrected sum of squares for $AB$ subclasses. Henderson (1959, 1969) has no formal definition of these terms. In some sense it is not necessary to give a formal definition because "everybody knows" what is meant. For example, the uncorrected sum of squares for the $AB$ subclasses is the sum over all such subclasses of the square of each subclass total, the sum being divided by the number of observations in such a subclass (the same number in each). However, Millman and Glass (1967) give a neat procedure for formalizing this. It starts from an expression for the total of all the observations. We state the rule using as an example the uncorrected sum of squares $bc$ in a situation where $x_{hijk}$ is the observation in levels $h$, $i$, $j$, and $k$ of factors $A$, $B$, $C$, and $W$, respectively.

**Rule 9.**

(a) Write down the total of all the observations obtaining

$$\sum_{h=1}^{n_a}\sum_{i=1}^{n_b}\sum_{j=1}^{n_c}\sum_{k=1}^{n_w} x_{hijk.}$$

(b) Re-order the summation signs so that those pertaining to the letters in the symbolic form of the uncorrected sum of squares of interest ($bc$ in this case) come first, and enclose the remainder of the sum in parenthesis obtaining

$$\sum_{i=1}^{n_b}\sum_{j=1}^{n_c}\left(\sum_{h=1}^{n_a}\sum_{k=1}^{n_w} x_{hijk}\right).$$

(c) Square the parenthesis and divide by the product of the $n$'s therein. The result is the required sum of squares. We have that

$$bc = \frac{\displaystyle\sum_{i=1}^{n_b}\sum_{j=1}^{n_c}\left(\sum_{h=1}^{n_a}\sum_{k=1}^{n_w} x_{hijk}\right)^2}{n_a n_w}.$$

As a workable rule, this is patently simple.

### c. Expected Values of Mean Squares, $E$(MS)

Mean squares are sums of squares divided by degrees of freedom. We obtain expected values of mean squares, which we denote by $E$(MS), by the following rules.

#### (*i*) *Completely Random Models*

**Rule 10.** Denote variances by $\sigma^2$ with appropriate subscripts. There will be as many $\sigma^2$'s with corresponding subscripts as there are lines in the analysis of variance table. The variance corresponding to the $W$-factor is the error variance: $\sigma^2_{w:abc} = \sigma^2_e$.

**Example 5 Illustration of Variance Notation** When there is an $AC$: $B$ interaction, there is a variance $\sigma^2_{ac:b}$. □

When $n_w = 1$, there is no $W$-line in the analysis of variance, although it may be appropriate to think of $\sigma^2_w$ as existing.

**Rule 11.** Whenever a $\sigma^2$ appears in any $E$(MS), its coefficient is the product of all $n$'s whose subscripts do not occur in the subscript of that $\sigma^2$.

**Example 6 Illustration of Rule 11** The coefficient of $\sigma^2_{ac:b}$ is $n_w$ when the factors are $A$, $B$, $C$: $B$, and $W$: $ABC$. □

This rule implies that the coefficient of $\sigma^2_{w:abc}$ is always unity.

**Rule 12.** Each $E$(MS) contains only those $\sigma^2$'s (with coefficients) whose subscripts include all letters pertaining to that of the MS.

**Example 7 Illustration of Rule 12** For the $AC$: $B$ line, $E[MS(AC: B)] = n_w\sigma^2_{ac:b} + \sigma^2_{w:abc}$.

The above examples of Rules 10–12 are part of the expected values that we show in Table 9.4. These are the expected values, under the random model, of the mean squares of the analysis of variance. □

***(ii) Fixed Effects and Mixed Models***

**Rule 13.** Treat the model as completely random, except that the $\sigma^2$ terms corresponding to fixed effects and interactions of fixed effects get changed into quadratic functions of these fixed effects. All other $\sigma^2$ terms remain including those pertaining to interactions of fixed and random effects.

**TABLE 9.4 Example of Rules 10–12: Expected Values Under the Random Model of Mean Squares of Table 9.3**

| | Variances (Rule 10) and Coefficients (Rule 11) | | | | | |
|---|---|---|---|---|---|---|
| Mean Square | $n_bn_cn_w\sigma^2_a$ | $n_an_cn_w\sigma^2_b$ | $n_an_wn_w\sigma^2_{c:b}$ | $n_cn_w\sigma^2_{ab}$ | $n_w\sigma^2_{ac:b}$ | $\sigma^2_{w:abc} = \sigma^2_e$ |
| | Terms included (Rule 12) | | | | | |
| MS($A$) | * | | | * | * | * |
| MS($B$) | | * | * | * | * | * |
| MS($C$: $B$) | | | * | | * | * |
| MS($AB$) | | | | * | * | * |
| MS($AC$: $B$) | | | | | * | * |
| MS($W$: $ABC$) | | | | | | * |

*denotes a $\sigma^2$ term that is included; for example, $n_bn_cn_w\sigma^2_a$ is part of $E$[MS($A$)]

Rule 13 is equivalent to that given by Henderson (1969) but differs from Henderson (1959), where it is stated that some $\sigma^2$ terms "disappear" from some of the expectations of mean squares. Explanation of this difference is included in the discussion of the two-way classification that now follows.

## 7. THE TWO-WAY CLASSIFICATION

Chapter 7 deals fully with the analysis of unbalanced data from the fixed-effect model of the two-way classification. We repeat the analysis of variance for balanced models in Table 9.5, using new symbols for the sums of squares. For example, SSA is the sum of squares for the $A$-factor (after $\mu$), with

$$\text{SSA} = R(\alpha|\mu) = R(\alpha|\mu, \beta) = bn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2,$$

as in Table 7.9. We now develop expected values of these sums of squares for the fixed, random and mixed models. We do this both as an illustration of the "brute

**TABLE 9.5 Analysis of Variance for a Two-Way Classification Interaction Model, with Balanced Data (see Table 7.9)**

| Source of Variation | d.f. | Sum of Squares |
|---|---|---|
| Mean | 1 | $\text{SSM} = N\bar{y}_{...}^2$ |
| $A$-factor | $a-1$ | $\text{SSA} = bn\sum_{i=1}^{a}(\hat{y}_{i...} - \hat{y}_{...})^2.$ |
| $B$-factor | $b-1$ | $\text{SSB} = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$ |
| $AB$ interaction | $(a-1)(b-1)$ | $\text{SSAB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$ |
| Residual Error | $ab(n-1)$ | $\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$ |
| Total | $N = abn$ | $\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2$ |

| Mean Squares |
|---|
| $\text{MSM} = \text{SSM}$ |
| $\text{MSA} = \dfrac{\text{SSA}}{(a-1)}$ |
| $\text{MSB} = \dfrac{\text{SSB}}{(b-1)}$ |
| $\text{MSAB} = \dfrac{\text{SSAB}}{(a-1)(b-1)}$ |
| $\text{MSE} = \dfrac{\text{SSE}}{ab(n-1)}$ |

force" method of deriving such expectations and for discussing certain aspects of the mixed model.

As in Chapter 7, the equation of the model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \tag{11}$$

with $i = 1, 2, \ldots, a$, $j = 1,2, \ldots, b$ and since we are considering balanced data, $k = 1, 2, \ldots, n$. To establish expected values of the sums of squares in Table 9.5, first write down the various means. They involve using

$$\bar{\alpha}_{\cdot} = \frac{\sum_{i=1}^{a} \alpha_i}{a}, \quad \bar{\beta}_{\cdot} = \frac{\sum_{j=1}^{b} \beta_j}{b}$$

and $\bar{\gamma}_{i\cdot}$, $\bar{\gamma}_{\cdot j}$, and $\bar{\gamma}_{\cdot\cdot}$ defined in the analogous manner. Hence from (11),

$$\begin{aligned} \bar{y}_{i\cdot\cdot} &= \mu + \alpha_i + \bar{\beta}_{\cdot} + \bar{\gamma}_{i\cdot} + \bar{e}_{i\cdot\cdot}, \\ \bar{y}_{\cdot j\cdot} &= \mu + \bar{\alpha}_{\cdot} + \beta_j + \bar{\gamma}_{\cdot j} + \bar{e}_{\cdot j\cdot}, \\ y_{ij\cdot} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \bar{e}_{ij\cdot}, \end{aligned} \tag{12}$$

and

$$\bar{y}_{\cdots} = \mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot} + \bar{e}_{\cdot\cdot}.$$

Substituting (11) and (12) into Table 9.5 gives

$$\begin{aligned} \text{SSM} &= N(\mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot})^2 \\ \text{SSA} &= bn \sum_{i=1}^{a} (\alpha_i - \bar{\alpha}_{\cdot} + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot} + \bar{e}_{i\cdot\cdot} - \bar{e}_{\cdots})^2, \\ \text{SSB} &= an \sum_{i=1}^{a} (\beta_j - \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot j} - \bar{\gamma}_{\cdot\cdot} + \bar{e}_{\cdot j\cdot} - \bar{e}_{\cdots})^2, \\ \text{SSAB} &= n \sum_{i=1}^{a} \sum_{j=1}^{b} (\gamma_{ij} - \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot j} + \bar{\gamma}_{\cdot\cdot} + \bar{e}_{ij\cdot} - \bar{e}_{i\cdot\cdot} - \bar{e}_{\cdot j\cdot} + \bar{e}_{\cdots})^2 \end{aligned} \tag{13}$$

and

$$\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (e_{ijk} - \bar{e}_{ij\cdot})^2.$$

Now, no matter what model we use, fixed, random or mixed, we take the error terms as having zero mean and variance $\sigma_e^2$ and being independent of one another. Furthermore, the expected value of the product of an error term with $\mu$, an $\alpha$, a $\beta$, ora $\gamma$ is zero. (If the effects are fixed the products have zero expectation because $E(e_{ijk}) = 0$ and,

when any of the effects are random products with the $e$-terms have zero expectation because of assuming independence.) Finally, expected values of squares and products of means of the $e$'s are such that, for example,

$$E(\bar{e}_{i..}^2) = \frac{\sigma_e^2}{bn},$$
$$E(\bar{e}_{i..}\bar{e}_{...}) = E(\bar{e}_{.j.}\bar{e}_{...}) = E(\bar{e}_{ij.}\bar{e}_{...}) = \frac{\sigma_e^2}{abn}$$

and

$$E(\bar{e}_{i..}\bar{e}_{.j.}) = \frac{\sigma_e^2}{abn}.$$

Hence for the terms in (13),

$$\begin{aligned} E(\bar{e}_{i..} - \bar{e}_{...})^2 &= \tfrac{(a-1)\sigma_e^2}{abn}, \\ E(\bar{e}_{.j.} - \bar{e}_{...})^2 &= \tfrac{(b-1)\sigma_e^2}{abn}, \\ E(\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...})^2 &= \tfrac{(a-1)(b-1)\sigma_e^2}{abn}, \end{aligned} \tag{14}$$

and

$$E(e_{ijk} - e_{ij.})^2 = \frac{(n-1)\sigma_e^2}{n}.$$

Consequently, on taking expected values of (13) and by dividing by degrees of freedom to convert them to mean squares we get

$$\begin{aligned} E(\text{MSM}) &= EN(\mu + \bar{\alpha}_{.} + \bar{\beta}_{.} + \bar{\gamma}_{..})^2 + \sigma_e^2 \\ E(\text{MSA}) &= \tfrac{bn}{a-1}\sum_{i=1}^{a} E(\alpha_i - \bar{\alpha}_{.} + \bar{\gamma}_{i.} - \bar{\gamma}_{..})^2 + \sigma_e^2, \\ E(\text{MSB}) &= \tfrac{an}{b-1}\sum_{i=1}^{a} E(\beta_j - \bar{\beta}_{.} + \bar{\gamma}_{.j} - \bar{\gamma}_{..})^2 + \sigma_e^2, \\ E(\text{MSAB}) &= \tfrac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} E(\gamma_{ij} - \bar{\gamma}_{i.} - \bar{\gamma}_{.j} + \bar{\gamma}_{..})^2 + \sigma_e^2, \end{aligned} \tag{15}$$

and

$$E(\text{MSE}) = \sigma_e^2.$$

These results hold whether the model is fixed, random, or mixed. Each model determines the consequence of the expectation operations shown in the right-hand side of (15).

**TABLE 9.6 Expected Mean Squares of a Two-Way Classification Interaction Model, with Balanced Data (see Table 9.5)**

| Fixed-Effects Model | |
|---|---|
| $E(\text{MSM}) = N(\mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot})^2$ | $+ \sigma_e^2$ |
| $E(\text{MSA}) = \dfrac{bn}{a-1}\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_{\cdot} + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot})^2$ | $+ \sigma_e^2$ |
| $E(\text{MSB}) = \dfrac{an}{b-1}\sum_{j=1}^{b}(\beta_j - \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot j} - \bar{\gamma}_{\cdot\cdot})^2$ | $+ \sigma_e^2$ |
| $E(\text{MSAB}) = \dfrac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\gamma_{ij} - \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot j} + \bar{\gamma}_{\cdot\cdot})^2$ | $+ \sigma_e^2$ |
| $E(\text{SSE}) =$ | $\sigma_e^2$ |

### a. The Fixed Effects Model

In the fixed-effects model, all the $\alpha$'s, $\beta$'s, and $\gamma$'s are fixed effects. Therefore, the expectation operations on the right-hand sides of (15) just involve dropping the $E$ symbol. We show the results in Table 9.6. Their derivation does not make use of the "usual restrictions" on the elements of the model.

Suppose we consider a model

$$y_{ijk} = \mu' + \alpha'_i + \beta'_j + \gamma'_{ij} + e_{ijk} \tag{16}$$

where the "usual restrictions" are part of the model. These restrictions are

$$\sum_{i=1}^{a}\alpha'_i = 0, \quad \sum_{j=1}^{b}\beta'_j = 0, \tag{17a}$$

and

$$\sum_{i=1}^{a}\gamma'_{ij} = 0, \quad \text{for all } j, \sum_{j=1}^{b}\gamma'_{ij} = 0, \quad \text{for all } i. \tag{17b}$$

Before using these restrictions, the expected mean squares will be those of Table 9.6 with primes on the $\alpha$'s, $\beta$'s, and $\gamma$'s. After using the restrictions in (17a) and (17b), the expectations reduce to those of Table 9.6 because (17a) and (17b) implies $\bar{\alpha}'_{\cdot} = 0, \bar{\beta}'_{\cdot} = 0, \bar{\gamma}'_{\cdot j} = 0$ for all $j$ and $\bar{\gamma}'_{i\cdot} = 0$, for all $i$.

We can show that the apparent difference between Tables 9.6 and 9.7 is just that apparent and not real. Suppose we rewrite the model as

$$y_{ijk} = \mu + e_{ijk} \tag{18}$$

$$= \bar{\mu}_{\cdot\cdot} + (\bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot}) + (\bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot}) + (\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot}) + e_{ij\cdot} \tag{19}$$

**TABLE 9.7 Expected Mean Squares of a Two-Way Classification Interaction Model, with Balanced Data. (See Table 9.5)**

| Fixed-effects model, that includes the restrictions $\sum_{i=1}^{a}\alpha_i' = 0 = \sum_{j=1}^{b}\beta_j' = \sum_{i=1}^{a}\gamma_{ij}'$ for all $j$, and $\sum_{j=1}^{b}\gamma_{ij}' = 0$ for all $i$. | |
|---|---|
| $E(\text{MSM}) = N\mu'^2$ | $+\sigma_e^2$ |
| $E(\text{MSA}) = \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i'^2$ | $+\sigma_e^2$ |
| $E(\text{MSB}) = \frac{an}{b-1}\sum_{j=1}^{b}\beta_j'^2$ | $+\sigma_e^2$ |
| $E(MSAB) = \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{i=1}^{b}(\gamma_{ij}'^2$ | $+\sigma_e^2$ |
| $E(\text{MSE}) =$ | $\sigma_e^2$ |

Then, on defining,

$$\mu' = \bar{\mu}_{..}, \alpha_i' = \bar{\mu}_{i.} - \bar{\mu}_{..}, \beta_j' = \bar{\mu}_{.j} - \bar{\mu}_{..}, \quad \text{and} \quad \gamma_{ij}' = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}. \tag{20}$$

Equation (19) is identical to (16). Furthermore, by their definition in (20), $\alpha_i', \beta_j'$, and $\gamma_{ij}'$ satisfy the constraint equations in (17). We have, for example, $\sum_{i=1}^{a}\alpha_i' = \sum_{i=1}^{a}(\bar{\mu}_i - \bar{\mu}_{..}) = 0$.

Therefore, the definitions in (20) are consistent with the expected mean squares of Table 9.7. As a result, we have, for example,

$$E(\text{MSA}) = \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i'^2 + \sigma_e^2.$$

However, observe that when comparing (18) and (11),

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{21}$$

As a result, with

$$\alpha_i' = \bar{\mu}_{i.} - \bar{\mu}_{..}$$

we have from (20),

$$\begin{aligned}\alpha_i' &= \mu + \alpha_i + \bar{\beta}_{.} + \bar{\gamma}_{i.} - (\mu + \bar{\alpha}_{.} + \bar{\beta}_{.} + \bar{\gamma}_{..}) \\ &= \alpha_i - \bar{\alpha}_{.} + \bar{\gamma}_{i.} - \bar{\gamma}_{..}.\end{aligned} \tag{22}$$

**TABLE 9.8 Expected Mean Squares of a Two-Way Classification Interaction Model with Balanced Data (see Table 9.5)**

| Random-Effects Model | | | | |
|---|---|---|---|---|
| $E(\text{MSM}) =$ | $abn\mu^2 +$ | $bn\sigma_\alpha^2 +$ | $an\sigma_\beta^2 + n\sigma_\gamma^2 +$ | $\sigma_e^2$ |
| $E(\text{MSA}) =$ | | $bn\sigma_\alpha^2$ | $+ n\sigma_\gamma^2 +$ | $\sigma_e^2$ |
| $E(\text{MSB}) =$ | | | $an\sigma_\beta^2 + n\sigma_\gamma^2 +$ | $\sigma_e^2$ |
| $E(\text{MSAB}) =$ | | | $n\sigma_\gamma^2 +$ | $\sigma_e^2$ |
| $E(\text{MSE}) =$ | | | | $\sigma_e^2$ |

Thus, $\sum_{i=1}^{n} \alpha_i'^2$ of $E$(MSA) in Table (9.7) has the same meaning as does

$$\sum_{i=1}^{a} (\alpha_i - \bar{\alpha}_{\cdot} + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot})^2$$

of $E$(MSA) in Table 9.6. Hence, interpretation of the $F$-statistic MSA/MSE is the same whether one uses Table 9.6 or 9.7. The $F$-statistic MSA/MSE tests the significance of $\alpha$-effects in the presence of (or, plus the average of) interaction effects. In Table 9.7 the symbols are defined, as in (17) so that these averages are zero whereas in Table 9.6, they are not so defined. We demonstrated the equivalence of the expressions for $E$(MSA) in Tables 9.6 and 9.7 via equation (22). In like manner, we can also demonstrate equivalence of other entries in the two tables by basing them on

$$\begin{aligned} \beta_j' &= \beta_j - \bar{\beta}_{\cdot} + \gamma_{\cdot j} - \bar{\gamma}_{\cdot\cdot}, \\ \gamma_{ij}' &= \gamma_{ij} - \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot j} + \bar{\gamma}_{\cdot\cdot} \end{aligned} \tag{23}$$

and

$$\mu' = \mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot}.$$

Defining effects that satisfy "the usual restrictions" in the manner of (20) results in the simplification of Table 9.6 to the form of Table 9.7. However, this simplification only occurs for balanced data. It does not occur for unbalanced data because sums of squares used with such data (e.g., Table 9.8) have expected values that do not involve the means of the effects in such a simple manner as with balanced data (see Table 9.6). Sometimes for unbalanced data, restrictions that are in terms of weighted sums of the effects are suggested. However, these have no simplifying effect when there are empty cells, as is often the case with unbalanced data.

A special case of the simplifying effect of the "usual restrictions" (20) that is of some interest is $E$(MSM). In Table 9.6

$$E(\text{MSM}) = N(\mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot}) + \sigma_e^2 = N[E(\bar{y})] + \sigma_e^2, \tag{24}$$

consistent with the hypothesis $H: E(\bar{y}) = 0$ that can be tested by the $F$-statistic. In Table 9.7, the expected value is

$$E(MSM) = N\mu'^2 + \sigma_e^2$$

consistent with testing, in *that* model $H: E(\bar{y}) = \mu' = 0$. This is the origin of the concept of "testing the mean" by the $F$-statistic $F(M) = \text{MSM/MSE.}$, referred to in earlier chapters. There, with unbalanced data, we saw how the meaning of this phrase was best described in terms of testing $H: E(\bar{y}) = 0$. For balanced data, that description is still appropriate when the model has "no usual restrictions," as is evident in (24). However, when the model does include such restrictions, the hypothesis $H: E(\bar{y}) = 0$ reduces to $H: \mu' = 0$ and thus gets described as "testing the mean."

### b. Random-Effects Model

All the $\alpha$-, $\beta$-, and $\gamma$-effects are random in the random-effects model, with zero means and variances $\sigma_\alpha^2, \sigma_\beta^2$, and $\sigma_\gamma^2$. Thus, for example,

$$E(\alpha_i) = 0 \quad \text{and} \quad E(\alpha_i^2) = \sigma_\alpha^2. \tag{25}$$

The effects are also assumed to be uncorrelated with each other. Hence,

$$E(\alpha_i \beta_j) = 0 = E(\alpha_i \gamma_{ij}) \quad \text{and} \quad E(\alpha_i \alpha_{i'}) = 0 \quad \text{for} \quad i \neq i'. \tag{26}$$

Furthermore, similar to (14),

$$E(\alpha_i - \bar{\alpha}_\cdot)^2 = \frac{(a-1)\sigma_\alpha^2}{a}. \tag{27}$$

Similar results hold for the $\beta$'s and the $\gamma$'s. Using them in (15) gives the expectations shown in Table 9.8.

Estimation of the variance components from Table 9.8 is achieved by equating mean squares to expected values, the resulting solutions for the components being the estimators. This gives

$$\begin{aligned} \hat{\sigma}_e^2 &= \text{MSE}, & \hat{\sigma}_\beta^2 &= \frac{(\text{MSB} - \text{MSAB})}{an}, \\ \hat{\sigma}_\gamma^2 &= \frac{(\text{MSAB} - \text{MSE})}{n}, & \hat{\sigma}_\alpha^2 &= \frac{(\text{MSA} - \text{MSAB})}{bn}. \end{aligned} \tag{24}$$

**Example 8 Analysis of a Two-Way Model with Both Factors Random** Houf and Berman (1988) describe an experiment conducted to investigate the capability of measurements on thermal impedance ($C^\circ/w \times 100$) on a power module for an

induction motor starter. There are 10 parts selected at random and three randomly selected inspectors. Each inspector takes three observations. The data follow:

| Inspector | 1 | | | 2 | | | 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Test | | | Test | | | Test | | |
| Part no. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 1 | 37 | 38 | 37 | 41 | 41 | 40 | 41 | 42 | 41 |
| 2 | 42 | 41 | 43 | 42 | 42 | 42 | 43 | 42 | 43 |
| 3 | 30 | 31 | 31 | 31 | 31 | 31 | 29 | 30 | 28 |
| 4 | 42 | 43 | 42 | 43 | 43 | 43 | 42 | 42 | 42 |
| 5 | 28 | 30 | 29 | 29 | 30 | 29 | 31 | 29 | 29 |
| 6 | 42 | 42 | 43 | 45 | 45 | 45 | 44 | 46 | 45 |
| 7 | 25 | 26 | 27 | 28 | 28 | 30 | 29 | 27 | 27 |
| 8 | 40 | 40 | 40 | 43 | 42 | 42 | 43 | 43 | 41 |
| 9 | 25 | 25 | 25 | 27 | 29 | 28 | 26 | 26 | 26 |
| 10 | 35 | 34 | 34 | 35 | 35 | 34 | 35 | 34 | 35 |

A two-way analysis of variance produced the following output using SAS.

| **Class Level Information** | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **part** | 10 | 1 2 3 4 5 6 7 8 9 10 |
| **inspector** | 3 | 1 2 3 |

| | |
|---|---|
| **Number of Observations Read** | 90 |
| **Number of Observations Used** | 90 |

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **Model** | 29 | 4023.733333 | 138.749425 | 271.47 | <.0001 |
| **Error** | 60 | 30.666667 | 0.511111 | | |
| **Corrected Total** | 89 | 4054.400000 | | | |

| **R-Square** | **Coeff Var** | **Root MSE** | **result Mean** |
|---|---|---|---|
| 0.992436 | 1.996984 | 0.714920 | 35.80000 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
|---|---|---|---|---|---|
| **part** | 9 | 3935.955556 | 437.328395 | 855.64 | <.0001 |
| **inspector** | 2 | 39.266667 | 19.633333 | 38.41 | <.0001 |
| **part*inspector** | 18 | 48.511111 | 2.695062 | 5.27 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **part** | 9 | 3935.955556 | 437.328395 | 855.64 | <.0001 |
| **inspector** | 2 | 39.266667 | 19.633333 | 38.41 | <.0001 |
| **part*inspector** | 18 | 48.511111 | 2.695062 | 5.27 | <.0001 |

| source | Type III Expected Mean Square |
|---|---|
| **part** | Var(Error) + 3 Var(part*inspector) + 9 Var(part) |
| **inspector** | Var(Error) + 3 Var(part*inspector) + 30 Var(inspector) |
| **part*inspector** | Var(Error) + 3 Var(part*inspector) |

| Tests of Hypotheses Using the Type III MS for part*inspector as an Error Term | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **part** | 9 | 3935.955556 | 437.328395 | 162.27 | <.0001 |
| **inspector** | 2 | 39.266667 | 19.633333 | 7.28 | 0.0048 |

First, observe from the $p$-values that part, inspector and interaction are significant at the 1% level. For evaluating significance of part and inspector, we use the interaction term as the denominator for the $F$-test instead of the error term. The reason for this will be made clear in Section 9c. In our notation based on the Type III expected mean square above, the system of equations for estimating the variance components by the analysis of variance method is

$$\begin{aligned} 9\hat{\sigma}_{\alpha}^2 + 3\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{e}^2 &= 437.33 \\ 30\hat{\sigma}_{\beta}^2 + 3\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{e}^2 &= 19.63 \\ 3\hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{e}^2 &= 2.70 \\ \hat{\sigma}_{e}^2 &= 0.51 \end{aligned}$$

Solving for the estimates of the variance components, we have,

$$\begin{aligned} \hat{\sigma}_{e}^2 &= 0.51 \\ \hat{\sigma}_{\gamma}^2 &= \frac{2.70 - 0.51}{3} = 0.73 \\ \hat{\sigma}_{\beta}^2 &= \frac{19.63 - 2.70}{30} = 0.560 \\ \hat{\sigma}_{\alpha}^2 &= \frac{437.33 - 2.70}{9} = 48.292. \end{aligned}$$

**TABLE 9.9 Expected Mean Squares of a Two-Way Classification Interaction Model with Balanced Data**

| Mixed Model $\alpha$'s Fixed, and $\beta$'s and $\gamma$'s Random | |
|---|---|
| $E(\text{MSM}) = abn(\mu + \bar{\alpha}_\cdot)^2$ | $+ an\sigma_\beta^2 + n\sigma_\gamma^2 + \sigma_e^2$ |
| $E(\text{MSA}) = \dfrac{bn}{a-1}\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_\cdot)^2$ | $+ n\sigma_\gamma^2 + \sigma_e^2$ |
| $E(\text{MSB}) =$ | $an\sigma_\beta^2 + n\sigma_\gamma^2 + \sigma_e^2$ |
| $E(\text{MSAB}) =$ | $n\sigma_\gamma^2 + \sigma_e^2$ |
| $E(\text{MSE}) =$ | $\sigma_e^2$ |

### c. The Mixed Model

Suppose the $\alpha$-effects are fixed effects and the $\beta$'s and $\gamma$'s are random. Then the expectation operations on the right-hand side of (15) involve dropping the $E$ symbol insofar as it pertains to the $\alpha$'s and using properties like those of (25), (26), and (27) for $\beta$'s and $\gamma$'s. This leads to the results shown in Table 9.9.

The difference between the random and mixed models is that the $\alpha$'s are random effects in the random model and are fixed effects in the mixed model. Since only the first two equations in (15) involve $\alpha$'s, only the first two entries in Table 9.9 differ from the corresponding entries in Table 9.8 and then only in having quadratic terms in the $\alpha$'s instead of terms in $\sigma_\alpha^2$.

The expectations in Table 9.9 are arrived at without making any use of the "usual restrictions" on elements of the model, just as are the expectations in Table 9.6 for the fixed-effects model. However, if the restriction $\sum_{i=1}^{a}\alpha_i = 0$ is taken as part of the mixed model, then, $E$(MSA) of Table 9.9 reduces to

$$E(\text{MSA}) = \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i^2 + n\sigma_\gamma^2 + \sigma_e^2,$$

the quadratic in the $\alpha$'s being similar to that of Table 9.8.

An often-used alternative mixed model is

$$y_{ijk} = \mu'' + \alpha_i'' + \beta_j'' + \gamma_{ij}'' + e_{ijk} \tag{29}$$

with the restriction

$$\sum_{i=1}^{a}\gamma_{ij}'' = \gamma_{\cdot j}'' = 0 \quad \text{for all } j. \tag{30}$$

In (29), the $\alpha''$'s are fixed effects and the $\beta''$'s and $\gamma''$'s are random effects with zero means and variances $\sigma_{\beta''}^2$ and $\sigma_{\gamma''}^2$, respectively. The $\beta''$'s and $\gamma''$'s are uncorrelated with each other and the $e$'s. Except for (30), this is all exactly the same as in the mixed model described earlier. This restriction implies a covariance between certain

of the $\gamma''$'s, namely between $\gamma''_{ij}$ and $\gamma''_{i'j}$ for $i \neq i'$. Suppose this covariance is the same,

$$\text{cov}(\gamma''_{ij}, \gamma''_{i'j}) = c \quad \text{for all} \quad i \neq i' \quad \text{and} \quad j. \tag{31}$$

Then from (30),

$$v\left(\sum_{i=1}^{a} \gamma''_{ij}\right) = 0.$$

Thus,

$$a\sigma^2_{\gamma''} + a(a-1)c = 0,$$

giving

$$c = -\frac{\sigma^2_{\gamma''}}{(a-1)}. \tag{32}$$

Note that this covariance pertains only to $\gamma''$'s within the same level of the $\beta$-factor, arising as it does from (30). The covariance between $\gamma''$'s in the same level of the $\alpha$-factor is zero as usual. Thus, we have

$$\text{cov}(\gamma''_{ij}, \gamma''_{ij'}) = 0 \quad \text{for all} \quad i \quad \text{and} \quad j \neq j'. \tag{33}$$

Prior to utilizing (30), we can derive the expected mean squares for the model (29) from equations (15) with double prime subscripts on $\mu$, the $\alpha$'s, $\beta$'s, and $\gamma$'s. Using $\bar{\gamma}''_{\cdot j} = 0$ from (30), we get that $\bar{\gamma}_{\cdot\cdot} = 0$. Then equations (15) become

$$\begin{aligned} E(\text{MSM}) &= N(\mu'' + \bar{\alpha}'')^2 + NE(\bar{\beta}''^2_{\cdot}) + \sigma^2_e, \\ E(\text{MSA}) &= \frac{bn}{a-1}\left[\sum_{i=1}^{a}(\alpha''_i - \bar{\alpha}''_{\cdot})^2 + \sum_{i=1}^{a} E(\bar{\gamma}''_{i\cdot})^2\right] + \sigma^2_e, \end{aligned} \tag{34a}$$

$$\begin{aligned} E(\text{MSB}) &= \frac{an}{b-1}\sum_{j=1}^{b} E(\beta''_j - \bar{\beta}''_{\cdot})^2 + \sigma^2_e, \\ E(\text{MSAB}) &= \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} E(\gamma''_{ij} - \bar{\gamma}''_{i\cdot})^2 + \sigma^2_e \end{aligned} \tag{34b}$$

and

$$E(\text{MSE}) = \sigma^2_e.$$

**TABLE 9.10 Expected Mean Squares of a Two-Way Classification Interaction Model, with Balanced Data**

| Mixed Model, with Restrictions on Interaction Effects $\gamma_{.j} = 0$ for all $j$. | | | |
|---|---|---|---|
| $E(\text{MSM}) =$ | $N(\mu'' + \bar{\alpha}_{\cdot}'')^2$ | $+ an\sigma^2_{\beta''}$ | $+ \sigma^2_e$ |
| $E(\text{MSA}) =$ | $\dfrac{bn}{a-1}\sum_{i=1}^{a}(\alpha_i'' - \bar{\alpha}_{\cdot}'')^2 +$ | $n\left(\dfrac{a}{a-1}\right)\sigma^2_{\gamma''}$ | $+ \sigma^2_e$ |
| $E(\text{MSB}) =$ | | $an\sigma^2_{\beta''}$ | $+ \sigma^2_e$ |
| $E(\text{MSAB}) =$ | | $n\left(\dfrac{a}{a-1}\right)\sigma^2_{\gamma''}$ | $+ \sigma^2_e$ |
| $E(\text{MSE}) =$ | | | $\sigma^2_e$ |

In carrying out the expectation operations in $E(\text{MSA})$ and $E(\text{MSAB})$, we make use of (31), (32), and (33) to obtain

$$E(\bar{\gamma}_{i\cdot}'')^2 = \sigma^2_{\gamma''}\left[\frac{1}{b} + \frac{b(b-1)0}{b^2}\right] = \frac{\sigma^2_{\gamma''}}{b}$$

and

$$E(\gamma_{ij}'' - \bar{\gamma}_{i\cdot}'')^2 = \sigma^2_{\gamma''}\left(1 + \frac{1}{b} - \frac{2}{b}\right) = \frac{(b-1)}{b}\sigma^2_{\gamma''}.$$

As a result, expressions (34) reduce to those of Table 9.10.

The results in Table 9.10 differ from those in Table 9.9 in two important ways. The expectations $E(\text{MSB})$ and $E(\text{MSM})$ do not contain $\sigma^2_{y''}$. Furthermore, the term in $\sigma^2_{y''}$ that does occur in $E(\text{MSA})$ and $E(\text{MSAB})$ includes the fraction $a/(a-1)$. The first of these differences, the absence of $\sigma^2_{y''}$ from, particularly $E(\text{MSB})$, is the reason for Rule 13 at the end of Section 6 differing from the first edition of Henderson (1959, 1969) but being the same as the second. The first edition specifies a general rule that leads to the absence of $\sigma^2_{y''}$ from $E(\text{MSB})$ on the basis of $\gamma_{\cdot j}'' = 0$ as in (30). The second edition specifies a general rule that retains $\sigma^2_{y''}$ in $E(\text{MSB})$ as in Table 9.9, using a model that has no restriction like (30). This dual approach to the mixed model is evident in many places. For example, Mood (1950, p. 344) and Kirk (1968, p. 137) use the Table 9.9 expectations. Anderson and Bancroft (1952, p. 339), Scheffe (1959, p. 269), Graybill (1961, p. 398), and Snedecor and Cochran (1967, p. 367) use those akin to Table 9.10. Mood and Graybill (1963) do not discuss the topic. Results like Table 9.10 predominate in the literature. However, Hartley and Rao (1967) point out that a strong argument for using the results of Table 9.9 is that they are consistent with the results for unbalanced data.

The second difference between Tables 9.9 and 9.10 is the occurrence of $a/(a-1)$ in terms of the interaction variance component in Table 9.10. This is a consequence of the restriction $\bar{\gamma}_{\cdot j}'' = 0$ of (30). Steel and Torrie (1960, pp. 214, 246), for example, also show this.

We now establish a relationship between Tables 9.9 and 9.10. The model for Table 9.9 is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

with the $\alpha$'s as fixed effects and the $\beta$'s and the $\gamma$'s random. We rewrite it as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \bar{\gamma}_{.j} + \gamma_{ij} - \bar{\gamma}_{.j} + e_{ijk}.$$

On defining $\mu'' = \mu, \alpha_i'' = \alpha_i$,

$$\beta_j'' = \beta_j + \overline{\gamma}_{.j} \quad \text{and} \quad \gamma_{ij}'' = \gamma_{ij} - \overline{\gamma}_{.j}, \tag{35}$$

we have the model (29), corresponding to Table 9.10. This follows because from (35),

$$\gamma_{.j}'' = \gamma_{.j} - \gamma_{.j} = 0$$

as in (30). Other properties of the $\gamma''$'s also follow. First, from (35),

$$\sigma_{\beta''}^2 = \sigma_\beta^2 + \frac{\sigma_\gamma^2}{a} \tag{36}$$

and

$$\sigma_{\gamma''}^2 = \sigma_\gamma^2 \left(1 + \frac{1}{a} - \frac{2}{a}\right) = \frac{(a-1)\sigma_\gamma^2}{a}, \tag{37}$$

giving

$$\sigma_\gamma^2 = \frac{a}{a-1}\sigma_{\gamma''}^2. \tag{38}$$

In addition, we have that

$$\begin{aligned}
\operatorname{cov}(\beta_j'', \gamma_{ij}'') &= \sigma_\gamma^2 \left(\frac{1}{a} - \frac{1}{a}\right) = 0, \\
\operatorname{cov}(\beta_j'', \gamma_{ij'}'') &= 0 \quad \text{for} \quad j \neq j', \\
\operatorname{cov}(\gamma_{ij}'', \gamma_{i'j}'') &= \sigma_\gamma^2 \left(-\frac{1}{a} - \frac{1}{a} + \frac{1}{a}\right) \\
&= -\frac{\sigma_\gamma^2}{a} \\
&= -\frac{\sigma_{\gamma''}^2}{(a-1)}
\end{aligned}$$

from (38). Furthermore,

$$\operatorname{cov}(\gamma''_{ij}, \gamma''_{ij'}) = 0$$

as in (33). Hence, the properties of the $\beta''$'s and $\gamma''$'s defined in (35) are exactly those attributed to (29) and (30) in deriving Table 9.10. In addition, substituting (36) and (38) into Table 9.10 yields Table 9.9[1].

Lengthy discussion of the models (29) and (30) that leads to Table 9.10 is available in Wilk and Kempthorne (1955, 1956), Cornfield and Tukey (1956), and Scheffe (1959). The model that leads to Table 9.9 is the one customarily used for unbalanced data. More than this will not be said. The purpose of this section was to show a relationship between the two different models. In either model, the variance components are estimated from the last three mean squares of the appropriate table, either 9.9 or 9.10.

**Example 9 Expected Mean Squares If We Assume Inspectors in Example 8 Are Fixed and Parts are Still Random** If in Example 8, we assume that inspectors are specific people but the parts are chosen at random, the expected means squares of Table 9.9 are given below. We have,

$$\begin{aligned}
E(\text{MS parts}) &= 9\sigma_\alpha^2 &&+ 3\sigma_\gamma^2 + \sigma_e^2,\\
E(\text{MS inspectors}) &= 15\sum_{j=1}^{3}(\beta_j - \bar{\beta}_\cdot)^2 &&+ 3\sigma_\gamma^2 + \sigma_e^2,\\
E(\text{MS parts*inspectors}) &= &&3\sigma_\gamma^2 + \sigma_e^2,\\
\text{and } E(\text{MSE}) &= &&\sigma_e^2
\end{aligned}$$

The reader may estimate the variance components in Exercise 5a. Using SAS, we get the following tables, the numerical results for the different sums of squares being the same.

**The SAS System**

**The GLM Procedure**

**Quadratic Forms of Fixed Effects in the Expected Mean Squares**

| **Source: Type III Mean Square for inspector** | | | |
|---|---|---|---|
| | **inspector 1** | **inspector 2** | **inspector 3** |
| **inspector 1** | 20.00000000 | −10.00000000 | −10.00000000 |
| **inspector 2** | −10.00000000 | 20.00000000 | −10.00000000 |
| **inspector 3** | −10.00000000 | −10.00000000 | 20.00000000 |

[1] S. R. Searle gratefully acknowledges conversations on this topic with C. R. Henderson, R. R. Hocking, and N. S. Urquhart.

**The SAS System**

**The GLM Procedure**

| Source | Type III Expected Mean Square |
|---|---|
| **part** | Var(Error) + 3 Var(part*inspector) + 9 Var(part) |
| **inspector** | Var(Error) + 3 Var(part*inspector) + Q(inspector) |
| **part*inspector** | Var(Error) + 3 Var(part*inspector) |

**The SAS System**

**The GLM Procedure**

**Dependent Variable: result**

| Tests of Hypotheses Using the Type III MS for part*inspector as an Error Term | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **part** | 9 | 3935.955556 | 437.328395 | 162.27 | <.0001 |
| **inspector** | 2 | 39.266667 | 19.633333 | 7.28 | 0.0048 |

The quadratic form uses the matrix given in the first line of the table divided by the degrees of freedom for inspectors, which is 2. Thus,

$$Q = \frac{1}{2}\begin{bmatrix}\beta_1 & \beta_2 & \beta_3\end{bmatrix}\begin{bmatrix}20 & -10 & -10\\ -10 & 20 & -10\\ -10 & -10 & 20\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}.$$

In Exercise 5b, the reader can show that $Q = 15\sum_{j=1}^{3}(\beta_j - \bar{\beta}_\cdot)^2$.

## 8. ESTIMATING VARIANCE COMPONENTS FROM BALANCED DATA

The method of estimating variance components from balanced data has been discussed and illustrated in terms of the one-way and two-way classifications. The rules of Section 6 determine both the appropriate analysis of variance and their expected mean squares. The expected mean squares are equated to the observed mean squares for obtaining estimators. We now discuss the properties of estimators derived in this fashion.

We now discuss the properties of estimators derived by the analysis of variance method.

For illustration, we use the one-way classification model. Table 9.11 shows its analysis of variance and is a generalization of Table 9.2. We consider data consisting of $a$ classes with $n$ observations in each class.

**TABLE 9.11 Analysis of Variance for One-Way Classification Random Model, $n$ Observations in Each of $a$ Classes $N = an$**

| Source of Variation | d.f. | Sum of Squares | Mean Square | Expected Value of Mean Square |
|---|---|---|---|---|
| Mean | 1 | $\text{SSM} = T_\mu$ | $\text{MSM} = \text{SSM}$ | $N\mu^2 + n\sigma_\alpha^2 + \sigma_e^2$ |
| Classes | $a - 1$ | $\text{SSA} = T_A - T_\mu$ | $\text{MSA} = \dfrac{\text{SSA}}{a-1}$ | $n\sigma_\alpha^2 + \sigma_e^2$ |
| Residual error | $a(n-1)$ | $\text{SSE} = T_0 - T_A$ | $\text{MSE} = \dfrac{\text{SSE}}{a(n-1)}$ | $\sigma_e^2$ |
| Total | $an$ | $\text{SST} = T_0$ | | |

We introduce and use the notation

$$T_o = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2, \qquad T_A = n\sum_{i=1}^{a} \bar{y}_{i.}^2, \quad \text{and} \quad T_\mu = N\bar{y}_{..}^2 \tag{39}$$

with $N = an$ in Table 9.11 because:

1. It refers to the basic calculations required;
2. It simplifies the writing of the analysis of variance tables; and
3. It extends conveniently to unbalanced data.

Each $T$-term is a total uncorrected sum of squares, with the subscript indicating the factor it refers to. The subscript *0* is for the observations. The subscript $A$ is for the $A$-factor. The subscript $\mu$ for $T_\mu = R(\mu)$.

Estimation of $\sigma_\alpha^2$ and $\sigma_e^2$ follows from Table 9.11 in the same way it follows from Table 9.2. Thus, we have that

$$\hat{\sigma}_e^2 = \text{MSE} \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{(\text{MSA} - \text{MSE})}{n}. \tag{40}$$

**Notation:** From now on, we abandon the use of ˆ over a symbol to denote best linear unbiased estimator. Instead ˆ will simply mean "an estimator of."

### a. Unbiasedness and Minimum Variance

Estimators of variance components obtained from balanced data are unbiased, regardless of whether the model is fixed or random. Suppose that $\mathbf{m} = \{M_i\}$, for $i = 1, 2, \ldots, k$, is the vector of mean squares such that $E(\mathbf{m})$ does not involve fixed effects. Furthermore, $\sigma^2$ is the vector of variance components to be estimated, with $E(\mathbf{m}) = \mathbf{P}\sigma^2$ for $\mathbf{P}$ non-singular. Then, we can solve the equations $\mathbf{m} = \mathbf{P}\sigma^2$ as

$$\hat{\sigma}^2 = \mathbf{P}^{-1}\mathbf{m} \tag{41}$$

**TABLE 9.12 Hypothetical Data of a One-Way Classification, Three Observations in Two Classes**

| Class | Observations | | | Total |
|---|---|---|---|---|
| 1 | 19 | 17 | 15 | $51 = y_{1.}$ |
| 2 | 25 | 5 | 15 | $45 = y_{2.}$ |
| | | | | $96 = y_{..}$ |

for the estimators of the variance components. These estimators are unbiased because

$$E(\hat{\sigma}^2) = \mathbf{P}^{-1}E(\mathbf{m}) = \mathbf{P}^{-1}\mathbf{P}\sigma^2 = \sigma^2.$$

The property of unbiasedness applies to both random and mixed models for balanced data only. The analogous estimators for unbalanced data are not unbiased. We return to this point in Chapter 10. Now, we merely note that even this simplest of properties, unbiasedness is not universally true for analysis of variance estimators of variance components.

The estimators of $\hat{\sigma}^2$ of (41) have the smallest variance of all estimators which are both quadratic functions of the observations and unbiased. Graybill and Hultquist (1961) present this property of minimum quadratic unbiasedness. Under normality assumptions, the estimators in (41) have the smallest variance from among all unbiased estimators, both those that are quadratic functions of the observations and those that are not. Graybill (1954) and Graybill and Wortham (1956) discuss this property. These papers, and the minimum variance properties they establish, apply only to balanced data. Discussion of similar properties for unbalanced data from the one-way classification is available in Townsend (1968) and Harville (1969a). Some discussion for the two-way classification can be found in Searle, Casella, and Mc Culloch (1992).

### b. Negative Estimates

A variance component is, by definition, positive. Nevertheless, estimates derived from (41) could be negative. A simple example illustrates this. Suppose three observations in each of two classes are those of Table 9.12.Then as in (39),

$$\begin{aligned} T_A &= \frac{51^2}{3} + \frac{45^2}{3} && = 1542 \\ T_\mu &= \frac{96^2}{6} && = 1536 \\ \mathrm{T}_0 &= 19^2 + 17^2 + 15^2 + 25^2 + 5^2 + 15^2 = 1750. \end{aligned}$$

Table 9.13 shows the analysis of variance for the data of Table 9.12. Hence, as in (40),

$$\hat{\sigma}_e^2 = 52 \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{6 - 52}{3} = -15.333. \tag{42}$$

**TABLE 9.13 Analysis of Variance in Data in Table 9.12**

| Source | d.f. | Sum of Squares | Mean Square | Expected Mean Square |
|---|---|---|---|---|
| Mean | 1 | $1536 = 1536$ | 1536 | |
| Classes | 1 | $1542 - 1536 = 6$ | 6 | $3\sigma_a^2 + \sigma_e^2$ |
| Residual error | 4 | $1750 - 1542 = 208$ | 52 | $\sigma_e^2$ |
| Total | 6 | $1750 = 1750$ | | |

This demonstrates how negative estimates can arise from the analysis of variance method. There is nothing intrinsic in the method to prevent this. This can happen not only in a simple case such as (42) but also in many factored models, both with balanced and unbalanced data.

It is clearly embarrassing to estimate a variance component as negative. Interpretation of a negative estimate of a non-negative parameter is obviously a problem. Several courses of action exist. Few are satisfactory. We list some possibilities.

(i) Accept the estimate, despite its distastefulness and use it as evidence that the true value of the component is zero. Although this interpretation may be appealing, the unsatisfying nature of the negative estimate still remains. This is particularly true if the negative estimate is used to estimate the sum of components. In that case, the estimated sum can be less than the estimate of an individual component. For example, from (42), we have the estimated sum of the components as $\hat{\sigma}_a^2 + \hat{\sigma}_e^2 = 52 - 15.333 = 36.667 < \hat{\sigma}_e^2$.

(ii) Accept the negative estimate as evidence that the true value of the corresponding component is zero. Hence, use zero in place of the negative value. Even though this seems to be a logical replacement, such a truncation procedure disturbs the properties of the estimates as otherwise obtained. For example, they are no longer unbiased.

(iii) Use the negative estimate as an indication of a zero component and ignore that component while retaining the factor as far as the lines in the analysis of variance table are concerned. This leads to ignoring the component estimated as negative and re-estimating the others. Thompson (1961, 1962) gives rules for doing this, known as "pooling minimal mean squares with predecessors." Thompson and Moore (1963) give an application.

(iv) Interpret the negative estimate as indication of a wrong model and re-examine the source of the data to look for a new model. Searle and Fawcett (1970) suggest finite population models as possible alternatives. They sometimes give positive estimates when infinite populations models yield negative estimates. However, their use is likely to be of limited extent. In contrast, Nelder (1954) suggests that at least for split plot and randomized block designs, randomization theory indicates that negative variance components can occur in some

situations. Such an apparent inconsistency can arise from the intra-block correlation of plots being less than the inter-block correlation.

(v) Use some method other than the analysis of variance method. There are several possibilities. One is to use Bayes procedures (see for example, Hill (1965, 1967), Tiao and Tan (1965, 1966), Tiao and Box (1967), Harville (1969b), and Searle, Casella, and Mc Culloch (1992). Another possibility is to use maximum likelihood estimators, as suggested by Herbach (1959) and Thompson (1962). We will discuss these estimators at the end of this chapter. Still another possibility is to use restricted maximum likelihood estimators. One might also try using the Minimum norm estimators (MINQUE) (see C. R. Rao (1970, 1971a,b) together with suggested modifications in Chaubey (1984). Two other references on the above-mentioned estimators are P. S. R. S. Rao (1997) and Searle (1995).

(vi) Take the negative estimate as indication of insufficient data. Follow the statistician's last hope. Collect more data and analyze them, either on their own or pooled with those that yielded the negative estimate. If the estimate from the pooled data were negative, that would be additional evidence that the corresponding component is zero.

Obtaining a negative estimate from the analysis of variance method is solely a consequence of the data and the method. It depends in no way on an implied distribution normality or otherwise. However, if we assume normality, it is possible to derive the probability of obtaining a negative estimate. We shall discuss this in Section 9e.

## 9. NORMALITY ASSUMPTIONS

Up to now, no particular form for the distribution of the error terms has been assumed. All of the preceding results obtained in this chapter are true for any distribution. We now make the normality assumptions. To be specific, we assume that the $e$'s and each set of random effects in the model are normally distributed, with zero means and variance–covariance structure discussed earlier. Recall that we assumed that the effects of each random factor have a variance–covariance matrix that is the variance component multiplied by an identity matrix and that the effects of each random factor are independent of those of every other factor and of the error terms. Under these conditions, we also assume normality.

### a. Distribution of Mean Squares

Let $f$, SS, and $M$ be the degrees of freedom, sum of squares and mean square

$$M = \frac{\text{SS}}{f}, \tag{43}$$

respectively, in a line of the analysis of variance table of *balanced data*. Under the normality assumptions just described, one can show that

$$\frac{\text{SS}}{E(M)} \sim \chi^2(f), \text{ and the SS-terms are pairwise independent.}$$

Hence,

$$\frac{fM}{E(M)} \sim \chi^2(f), \text{ and the } M's \text{ are pairwise independent.} \tag{44}$$

We can derive result (44) by writing SS/$E(M)$ as a quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ in the observations $\mathbf{y}$, and applying Theorems 5 and 6 of Chapter 2. In applying these theorems to random or mixed models, $\mathbf{V}$ is not $\sigma_e^2\mathbf{I}$ as it is in the fixed model. Instead, it a matrix whose elements are functions of the $\sigma^2$'s of the model as illustrated in (9) and (10). Nevertheless, for the $\mathbf{A}$-matrices in the quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ for SS/$E(M)$, we will find that $\mathbf{AV}$ is always idempotent. Furthermore, for the random model, $\boldsymbol{\mu}$ has the form $\mu\mathbf{1}$ and $\boldsymbol{\mu}'\mathbf{A}\mu = \mu'\mathbf{1}'\mathbf{A}\mathbf{1}\mu$ will, by the nature of $\mathbf{A}$, always be zero. Hence, the $\chi^2$'s are central, as indicated in (44). For the mixed model, (44) will also apply for all sums of squares whose expected values do not involve fixed effects. Those that do involve fixed effects will be non-central $\chi^2$'s.

**Example 10 The Distribution of the Between Sum of Squares in One-Way Analysis of Variance** The variance–covariance matrix for the one-way classification model of Table 9.11 is

$$\mathbf{V} = \sigma_e^2\mathbf{I} + \sigma_\alpha^2 \sum_{i=1}^{a} {}^{+}\mathbf{J} \tag{45}$$

where $\mathbf{I}$ has order $N = an$ and $\mathbf{J}$ has order $n$. The expression in (45) is a generalization of that in (10). Now for Table 9.11, with $\mathbf{J}_N$ being a $\mathbf{J}$-matrix of order $N$, the terms of (44) are

$$\text{SS} = \text{SSA} = \mathbf{y}'\left(n^{-1}\sum_{i=1}^{a} {}^{+}\mathbf{J} - N^{-1}\mathbf{J}_N\right)\mathbf{y} \tag{46}$$

and

$$E(M) = E(MSA) = n\sigma_a^2 + \sigma_e^2,$$

so that

$$\frac{\text{SS}}{E(M)} = \mathbf{y}'\mathbf{A}\mathbf{y} \quad \text{with} \quad \mathbf{A} = \frac{n^{-1}\sum_{i=1}^{a} {}^{+}\mathbf{J} - N^{-1}\mathbf{J}_N}{n\sigma_a^2 + \sigma_e^2}. \tag{47}$$

Hence, using properties of $\mathbf{J}$ matrices such as $\mathbf{1}'\mathbf{J} = n\mathbf{1}'$ and $\mathbf{J}^2 = n\mathbf{J}$ (see, for example, Searle (1966, p. 197) or Gruber (2014, p. 47)),

$$\begin{aligned}\mathbf{AV} &= \frac{\left[\sigma_e^2\left(n^{-1}\sum_{i=1}^{a} {}^{+}\mathbf{J} - N^{-1}\mathbf{J}_N\right) + \sigma_a^2\left(\sum_{i=1}^{a} {}^{+}\mathbf{J} - nN^{-1}\mathbf{J}_N\right)\right]}{n\sigma_a^2 + \sigma_e^2} \\ &= \sum_{i=1}^{a} {}^{+}n^{-1}\mathbf{J} - N^{-1}\mathbf{J}_N.\end{aligned} \tag{48}$$

It may be shown that $(\mathbf{AV})^2 = \mathbf{AV}$, meaning that $\mathbf{AV}$ is idempotent. Furthermore, from (47), $\mathbf{1}'\mathbf{A} = \mathbf{0}$. Hence,

$$\frac{\text{SSA}}{E(\text{MSA})} \sim \chi^{2'}[r(\mathbf{AV}), 0] = \chi^2(a-1). \tag{49}$$

The rank of $\mathbf{AV}$ is its trace, namely $a - 1$. This follows from (48).

There are, of course, easier ways to derive (49). However, the intermediary steps (45)–(48) have useful generalizations for the case of unbalanced data.

### b. Distribution of Estimators

Equating mean squares to their expected values as a method of deriving variance component estimators gives estimators that are linear functions of their mean squares. The distribution of these mean squares was given in (44). Therefore, the resulting variance components are linear functions of multiples of $\chi^2$-variables. Some of these $\chi^2$-variables have negative coefficients. A closed form does not exist for the distribution of such functions. Furthermore, the coefficients are themselves functions of the population variance components.

**Example 11 Demonstration That the Exact Form of the Variance Component $\sigma_\alpha^2$ cannot be Obtained** In Table 9.11, we have that

$$\frac{(a-1)\text{MSA}}{n\sigma_\alpha^2 + \sigma_e^2} \sim \chi^2(a-1).$$

Independently,

$$\frac{a(n-1)\text{MSE}}{\sigma_e^2} \sim \chi^2(an-a).$$

Therefore,

$$\begin{aligned}\hat{\sigma}_\alpha^2 &= \frac{\text{MSA} - \text{MSE}}{n} \\ &\sim \frac{n\sigma_\alpha^2 + \sigma_e^2}{n(a-1)}\chi^2(a-1) - \frac{\sigma_e^2}{an(n-1)}\chi^2(an-a).\end{aligned} \tag{50}$$

The exact form of the distribution in (50) cannot be derived for two reasons. First, its second term is negative. Second, $\sigma_\alpha^2$ and $\sigma_e^2$ occur in the coefficients and are unknown. □

In general, the exact form of the distribution cannot be derived for the type of variance component in Example 11 above. If we knew the coefficients of the variance components, we could employ the methods of Robinson (1965) or Wang (1967) to obtain their distributions as infinite series expansions.

In contrast to other components, we can obtain the exact form of the distribution of $\hat{\sigma}_e^2$ exactly, under normality assumptions. We have that

$$\hat{\sigma}_e^2 = \text{MSE} = \frac{\sigma_e^2}{f_{\text{MSE}}}\chi^2(f_{\text{MSE}}), \tag{51}$$

where $f_{\text{MSE}}$ are the degrees of freedom associated with MSE.

Generalization of (50) arises from (41), which is $\hat{\sigma}^2 = \mathbf{P}^{-1}\mathbf{m}$. The elements of $\mathbf{m}$ follow (44). Thus, for example, $M_i \sim E(M_i)f_i^{-1}\chi^2(f_i)$. Now, write

$$\mathbf{C} = \text{diag}\{f_i^{-1}\chi^2(f_i)\} \quad \text{for} \quad i = 1, 2, \ldots, k,$$

where there are $k$ lines in the analysis of variance being used. Then from (41),

$$\hat{\sigma}^2 \sim \mathbf{P}^{-1}\mathbf{C}E(\mathbf{m}) \sim \mathbf{P}^{-1}\mathbf{C}\mathbf{P}\sigma^2. \tag{52}$$

In this way, we can express the vector of estimators of variance components as a vector of multiples of central $\chi^2$ variables.

### c. Tests of Hypothesis

Expected values of mean squares, derived by the rules of Section 6 will suggest which mean squares are the appropriate denominators for testing the hypotheses that certain variance components are zero. In order to use a central $F$-distribution, the expected mean square of the numerator should be that of the denominator plus the variance component we wish to determine whether or not is significantly different from zero. Thus, in Table 9.9 MSAB/MSE is appropriate for testing the hypothesis $H\colon \sigma_\gamma^2 = 0$ and MSB/MSAB is appropriate for testing $H\colon \sigma_\beta^2 = 0$. Likewise, in Examples 8 and 9, we tested the hypotheses about the variance components for the main random effects

using the mean square for the interaction term. In the random model, all ratios of mean squares have central $F$-distributions, because all mean squares follow (44). In the mixed model, the same is true of ratios of mean squares whose expected values contain no fixed effects.

The table of expected values will not always suggest the "obvious" denominator for testing a hypothesis. For example, suppose in Table 9.4 we wish to test the hypothesis $\sigma_b^2 = 0$. From that table we have, using $M_1, M_2, M_2,$ and $M_4$, respectively for MS($B$), MS($C$: $B$), MS($AB$), and MS($AC$: $B$),

$$\begin{aligned}
E(M_1) &= k_1\sigma_b^2 + k_2\sigma_{c:b}^2 + k_3\sigma_{ab}^2 + k_4\sigma_{ac:b}^2 + \sigma_e^2 \\
E(M_2) &= \qquad k_2\sigma_{c:b}^2 \qquad\quad + k_4\sigma_{ac:b}^2 + \sigma_e^2 \\
E(M_3) &= \qquad\qquad\quad k_3\sigma_{ab}^2 + k_4\sigma_{ac:b}^2 + \sigma_e^2 \\
E(M_4) &= \qquad\qquad\qquad\quad k_4\sigma_{ac:b}^2 \quad + \sigma_e^2.
\end{aligned}$$

Here, we have written the coefficients of the $\sigma^2$'s as products of the $n$'s shown in the column heading of Table 9.4 as $k$'s, for example, $k_1 = n_a n_c n_w$. We observe from these expected values that there is no mean square in the table suitable for testing the hypotheses $H$: $\sigma_b^2 = 0$. The reason is that there is no mean square whose expected value is $E(M_1)$ with the $\sigma_b^2$ omitted. We see that

$$E(M_1) - k_1\sigma_b^2 = k_{\cdot 2}\sigma_{c:b}^2 + k_3\sigma_{ab}^2 + k_4\sigma_{ac:b}^2 + \sigma_e^2. \tag{53}$$

However, there is a linear function of the other means squares whose expected value does equal $E(M_1) - k_1\sigma_b^2$. We have that

$$E(M_2) + E(M_3) - E(M_4) = k_{\cdot 2}\sigma_{c:b}^2 + k_3\sigma_{ab}^2 + k_4\sigma_{ac:b}^2 + \sigma_e^2. \tag{54}$$

We shall show how to use the mean squares in (53) and (54) to calculate a ratio that is approximately distributed as a central $F$-distribution.

In (54), some of the mean squares are involved negatively. However, from (53), we have that

$$E(M_1) + E(M_4) = k_1\sigma_b^2 + E(M_2) + E(M_3).$$

Let us generalize this to

$$E(M_r + \cdots + M_s) = k\sigma_\alpha^2 + E(M_m + \cdots + M_n). \tag{55}$$

Consider testing the hypothesis $H$: $\sigma_\alpha^2 = 0$, where $\sigma_\alpha^2$ is any component of the model. To test this hypothesis, Satterthwaite (1946) suggests the statistic

$$F = \frac{M'}{M''} = \frac{M_r + \cdots + M_s}{M_m + \cdots + M_n}, \quad \text{which is approximately } \sim F(p, q) \tag{56}$$

where

$$p = \frac{(M_r + \cdots + M_s)^2}{M_r^2/f_r + \cdots + M_s^2/f_s} \quad \text{and} \quad q = \frac{(M_m + \cdots + M_n)^2}{M_m^2/f_m + \cdots + M_n^2/f_n}. \tag{57}$$

In $p$ and $q$, the term $f_i$ is the degrees of freedom associated with the mean square $M_i$.

The rationale for this test is that both the numerator and denominator of (56) are distributed approximately as multiples of central $\chi^2$ variables. each mean square in the analysis is distributed as the multiple of a central $\chi^2$. Moreover, in (56), there is no mean square that occurs in both the numerator and denominator. Thus, the random variables in the numerator and denominator of (56) are independent. Thus, we have that the statistic $F$ in (56) has an approximate $F(p, q)$-distribution.

In (56), both $M'$ and $M''$ are sums of mean squares. As Satterthwaite (1946) showed, $pM'/E(M')$ has an approximate central $\chi^2$-distribution with $p$ degrees of freedom where $p$ is given by (57). A similar result holds true for $M''$ with $q$ degrees of freedom. More generally, consider the case where some mean squares are included negatively. Suppose

$$M_0 = M_1 - M_2$$

where $M_1$ and $M_2$ are now *sums* of mean squares having degrees of freedom $f_1$ and $f_2$, respectively. Let

$$\rho = \frac{E(M_1)}{E(M_2)} \quad \text{and} \quad \hat{\rho} = \frac{M_1}{M_2} \geq 1.$$

In addition, let

$$\hat{f}_0 = \frac{(\hat{\rho} - 1)^2}{(\hat{\rho}/f_1 + 1/f_2)}.$$

Simulation studies of Gaylor and Hopper (1969) suggest that the statistic

$$\frac{\hat{f}_0 M_0}{E(M_0)} \text{ is approximately} \sim \chi^2(f_0)$$

provided that

$$\rho > F_{f_2 \cdot f_1 \cdot 0.975}, \quad f_1 \leq 100, \quad \text{and} \quad f_1 \leq 2f_2.$$

They further suggest that $\rho > F_{f_2 \cdot f_1 \cdot 0.975}$ "appears to be fulfilled reasonably well" when

$$\hat{\rho} > F_{f_2 \cdot f_1 \cdot 0.975} \times F_{f_2 \cdot f_1 \cdot 0.50}.$$

Under these conditions, we may use Satterthwaite's procedure in (56) and (57) on functions of mean squares that involve differences as well as sums.

### d. Confidence Intervals

Although we are unable to derive exact distributions, we can still find approximate and, in some cases, confidence intervals for functions of variance components.

Graybill (1961, p. 369) presents a method for obtaining approximate confidence intervals for a linear function of expected mean squares. The method is as follows. Define $\chi^2_{n,L}$ and $\chi^2_{n,U}$ as lower and upper points of a $(1-\alpha)\%$ region of the $\chi^2(n)$-distribution such that

$$\Pr\{\chi^2_{n,L} \le \chi^2(n) \le \chi^2_{n,U}\} = 1-\alpha. \tag{58}$$

Then for any constants $k_i$, such that $\sum k_i M_i > 0$, the approximate confidence interval on $\sum k_i M_i$ is given by

$$\Pr\left\{\frac{n\sum k_i M_i}{\chi^2_{r,U}} \le \sum k_i E(M_i) \le \frac{n\sum k_i M_i}{\chi^2_{r,L}}\right\} = 1-\alpha,$$

where

$$r = \frac{\left(\sum k_i M_i\right)^2}{\sum k_i^2 M_i^2/f_i}$$

analogous to (57). Since $r$ will seldom be an integer, $\chi^2_{r,L}$ and $\chi^2_{r,U}$ are obtained from tables of the central $\chi^2$-distribution using either interpolation or the nearest (or next largest integer to $r$. Welch gives a correction to the tabulated $\chi^2$-values when $r < 30$. Graybill (1961, p. 370) recommends its use and provides details. Other methods of finding simultaneous confidence intervals on ratios of variance components are available in Broemeling (1969).

Suppose that $M_1$ and $M_2$ are two mean squares having the properties of (44) and such that

$$E(M_1) = \theta + \sigma_e^2 \quad \text{and} \quad E(M_2) = \sigma_e^2.$$

Suppose $f_1$ and $f_2$ are the respective degrees of freedom of $M_1$ and $M_2$ and let

$$F = \frac{M_1}{M_2}.$$

Then with $F_{f_1,f_2,\alpha}$ being the upper $\alpha\%$ point of the $F(f_1,f_2)$-distribution, that is, a fraction $\alpha\%$ of the distribution lying beyond $F_{f_1,f_2,\alpha}$, write

$$\alpha_1 + \alpha_2 = \alpha$$

and

$$\begin{aligned} F_1 &= F_{f_2,f_1,\alpha}, \quad F_2 = F_{f_1,f_2,\alpha}, \\ F_1' &= F_{\infty,f_1,\alpha}, \quad F_2' = F_{f_1,\infty,\alpha}. \end{aligned}$$

Scheffe (1959, p. 235) gives an approximate $(1-\alpha)\%$ confidence interval that is similar to that of Bulmer (1957). The confidence interval on $\theta$ is

$$\left( \frac{M_2(F-F_2)(F+F_2-F_2')}{FF_2'}, \frac{M_2(F-1/F_1)(F+1/F_1-1/F_1')}{F/F_1'} \right).$$

When $F < F_2$, the lower limit is taken as zero. When $F < 1/F_1$, the interval is taken as zero.

Scheffe (1959, p. 235) (Also see Section 2 of Chapter 10 of that reference) indicates that this interval can be "seriously invalidated by non-normality, especially of the random effects" for which $M_1$ is the mean square.

Although, in general, only approximate confidence intervals can be placed on variance components, there are some instances where it is possible to derive exact intervals. The most notable is the interval for $\sigma_e^2$ based on the $\chi^2$-distribution of (51).

It yields the interval contained in the probability statement

$$\Pr\left\{ \frac{\text{SSE}}{\chi^2_{f_{\text{SSE}},U}} \le \sigma_e^2 \le \frac{\text{SSE}}{\chi^2_{f_{\text{SSE}},L}} \right\} = 1-\alpha \tag{59}$$

where for the degrees of freedom $f_{\text{SSE}} = f_{\text{MSE}}$. We derive the $\chi^2$-values from tables as in (58).

Other exact confidence intervals readily available are those for the one-way classification. We show these in Table 9.14 above. The first entry there is the appropriate form of (59). The last three entries are equivalent intervals for different ratio functions, all based on the fact that for $F = \text{MSA/MSE}$,

$$\frac{\sigma_e^2 F}{\left(n\sigma_\alpha^2 + \sigma_e^2\right)} \sim F[a-1,(n-1)]. \tag{60}$$

Graybill (1961, p. 379) gives the interval for $\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_e^2)$. Sheffe (1959, p. 229) gives the interval for $\sigma_\alpha^2/\sigma_e^2$. Williams (1962) gives the confidence interval for $\sigma_\alpha^2$, the second entry in the table. It results from combining (60) and the distribution of $\text{SSE}/\sigma^2$.

**TABLE 9.14 Confidence Intervals for the Variance Components and Functions Thereof, in the One-Way Classification Random Model, Balanced Data (See Table 9.11)**

| Parameter | Exact Confidence Interval[a] Lower limit | Exact Confidence Interval[a] Upper limit | Confidence Coefficient |
|---|---|---|---|
| $\sigma_e^2$ | $\dfrac{\text{SSE}}{\chi^2_{a(n-1),U}}$ | $\dfrac{\text{SSE}}{\chi^2_{a(n-1)L}}$ | $1-\alpha$ |
| $\sigma_\alpha^2$ | $\dfrac{\text{SSA}(1-F_U/F)}{n\chi^2_{a-1,U}}$ | $\dfrac{\text{SSA}(1-F_L/F)}{n\chi^2_{a-1,L}}$ | $1-2\alpha$ |
| $\dfrac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_e^2}$ | $\dfrac{F/F_U-1}{n+F/F_U-1}$ | $\dfrac{F/F_L-1}{n+F/F_L-1}$ | $1-\alpha$ |
| $\dfrac{\sigma_e^2}{\sigma_\alpha^2+\sigma_e^2}$ | $\dfrac{n}{n+F/F_L-1}$ | $\dfrac{n}{n+F/F_U-1}$ | $1-\alpha$ |
| $\dfrac{\sigma_\alpha^2}{\sigma_e^2}$ | $\dfrac{F/F_U-1}{n}$ | $\dfrac{F/F_L-1}{n}$ | $1-\alpha$ |

[a]Notation

$$F = \text{MSA/MSE}$$
$$\Pr\left\{\chi^2_{n,L} \le \chi^2(n) \le \chi^2_{n,U}\right\} = 1-\alpha$$
$$\Pr\{F_L \le F[a-1, a(n-1)] \le F_U\} = 1-\alpha.$$

### e. Probability of Negative Estimates

Consider two mean squares $M_1$ and $M_2$ of the kind described in (44). Suppose $E(M_1 - M_2) = k\sigma^2$ so that

$$\hat{\sigma}^2 = \frac{(M_1 - M_2)}{k}.$$

Then the probability of $\hat{\sigma}^2$ being negative is

$$\begin{aligned}\Pr\{\hat{\sigma}^2 \text{ is negative}\} &= \Pr\left\{\frac{M_1}{M_2} < 1\right\} \\ &= \Pr\left\{\frac{M_1/E(M_1)}{M_2/E(M_2)} < \frac{E(M_2)}{E(M_1)}\right\} \\ &= \Pr\left\{F(f_1,f_2) < \frac{E(M_2)}{E(M_1)}\right\}. \qquad (61)\end{aligned}$$

This provides a means of calculating the probability that an estimator of the form $\hat{\sigma}^2 = (M_1 - M_2)/k$ will be negative. It requires giving values to the variance components being estimated because $E(M_1)$ and $E(M_2)$ are functions of the components. However, in using a series of arbitrary values for these components, calculation of (61) provides some general indication of the probability of obtaining a negative estimate. The development of this procedure is due to Leone et al. (1968). We could extend it to

use the approximate $F$-statistic of (56) for finding the probability that the estimate of $\sigma_\alpha^2$ would be negative.

**Example 12 Probability that $\sigma_\alpha^2 < 0$ in the One-Way Classification** For the one-way classification of Table 9.11, equation (61) is

$$\Pr\left\{\hat{\sigma}_\alpha^2 < 0\right\} = \Pr\left\{F_{a-1,a(n-1)} < \frac{\sigma_e^2}{\sigma_e^2 + n\sigma_a^2}\right\}$$
$$= \Pr\left\{F_{a-1,a(n-1)} < \frac{1}{1+n\rho}\right\}$$

where $\rho = \sigma_\alpha^2/\sigma_e^2$. □

### f. Sampling Variances of Estimators

In spite of the fact that, in general, the distribution functions for variance component estimators that are linear functions of $\chi^2$-variables cannot be derived, their sampling variances can be. We shall show how to do this in this section. Of course, the variances are functions of the unknown components.

**(*i*) *Derivation.*** Since the estimators are linear functions of mean squares, they are linear functions of the quadratic forms of the observations. Hence, the estimators are also quadratic forms of the observations. Therefore, we may use Theorem 4 of Chapter 2 to derive their variances. We shall use this procedure for unbalanced data in Chapter 10. However, for balanced data, the mean squares are independent with known distributions, as in (44). Therefore, we can easily derive the variances of linear functions of the mean squares. If we write an estimator in the form

$$\hat{\sigma}^2 = \sum k_i M_i,$$

from (44), we have that $\text{cov}(M_i M_{i'}) = 0$ for $i \neq i'$ and

$$v(M_i) = 2f_i\left[\frac{E(M_i)}{f_i}\right]^2 = \frac{2[E(M_i)]^2}{f_i}.$$

Hence,

$$v(\hat{\sigma}^2) = 2\sum \frac{k_i^2[E(M_i)]^2}{f_i}. \tag{62}$$

**Example 13 Sampling Variance of Variance Components in the One-Way Classification** In the one-way classification of Table 9.11

$$\hat{\sigma}^2 = \frac{(\text{MSA} - \text{MSE})}{n}$$

and so from (62),

$$v(\hat{\sigma}_\alpha^2) = \frac{2}{n^2}\left[\frac{(n\sigma_\alpha^2 + \sigma_e^2)^2}{a-1} + \frac{\sigma_e^4}{a(n-1)}\right]. \tag{63}$$

Similarly from (51),

$$v(\sigma_e^2) = \frac{2\sigma_e^4}{f_{\text{MSE}}}. \tag{64}$$

For Table 9.11, this is

$$v(\sigma_e^2) = \frac{2\sigma_e^4}{a(n-1)}. \tag{65}$$

□

**(*ii*) *Covariance Matrix*** As noted in (44), mean squares in the analysis of variance are distributed independently of one another. Therefore, they have zero covariances. However, this is not necessarily the case for variance component estimators that are linear functions of these mean squares. Such estimators usually have non-zero covariances. For example, from (40), we have in the one-way classification

$$\text{cov}(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2) = -\frac{v(\text{MSE})}{n} \tag{66}$$

$$= -\frac{2\sigma_e^4}{an(n-1)}. \tag{67}$$

In general, from (41), the variance–covariance matrix of the vector of estimators is

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\text{var}(\mathbf{m})\mathbf{P}^{-1'}. \tag{68}$$

As a result of the mean squares being independent, var(**m**) is diagonal. Thus, we can write it as

$$\text{var}(\mathbf{m}) = \mathbf{D} = \text{diag}\left\{\frac{2[E(M_i)]^2}{f_i}\right\} \quad \text{for} \quad i = 1, 2, \ldots, k. \tag{69}$$

Then,

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\mathbf{D}\mathbf{P}^{-1'}. \tag{70}$$

Each element of (70) is a quadratic function of variance components, as is $[E(M_i)]^2$ in (69).

***(iii) Unbiased Estimation*** The estimation of the elements of $\text{var}(\hat{\sigma}^2)$ in any optimal manner is not easy because of the quadratic nature of the variance components. The procedure that is simplest and most often used is that of replacing $E(M_i)$ in $\mathbf{D}$ by $M_i$. Thus, from (69), we write

$$\mathbf{D}_1 = \text{diag}\left\{\frac{2M_i^2}{f_i}\right\} \quad \text{for} \quad i = 1, 2, \ldots, k. \tag{71}$$

We then have,

$$\tilde{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\mathbf{D}_1\mathbf{P}^{-1'}. \tag{72}$$

These estimators have no desirable properties. They are not even unbiased.

However, we can readily obtain unbiased estimators of $\text{var}(\hat{\sigma}^2)$ from (71) through replacing $f_i$ therein by $f_i + 2$. Thus, with

$$\mathbf{D}_2 = \text{diag}\left\{\frac{2M_i^2}{(f_i + 2)}\right\} \quad \text{for} \quad i = 1, 2, \ldots, k. \tag{73}$$

we have,

$$\hat{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\mathbf{D}_2\mathbf{P}^{-1'} \tag{74}$$

as an unbiased estimator of $\text{var}(\hat{\sigma}^2)$. For example, from (63) and (65),

$$\hat{v}(\hat{\sigma}_\alpha^2) = \frac{2}{n^2}\left[\frac{(n\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2)^2}{a+1} + \frac{\hat{\sigma}_e^4}{a(n-1)+2}\right]$$

and

$$\hat{v}(\hat{\sigma}_e^2) = \frac{2\hat{\sigma}_e^4}{a(n-1)+2}$$

are unbiased estimators of the variances $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$, respectively.

The reason that (74) gives an unbiased estimator of $\text{var}(\hat{\sigma}^2)$ is as follows. For any mean square $M$, with degrees of freedom $f$,

$$v(M) = \frac{2[E(M)]^2}{f}.$$

By the well-known shortcut formula for calculating the variance of a random variable,

$$v(M) = E(M^2) - [E(M)]^2.$$

Hence,

$$E(M^2) = \left(1 + \frac{2}{f}\right)[E(M)]^2.$$

As a result, $M^2/(f+2)$ is an unbiased estimator of $[E(M)]^2/f$. Therefore, using $M_i^2/(f_i+2)$ in place of $E(M_i^2)/f_i$ in (69) as is done in (73), makes $\mathbf{D}_2$ an unbiased estimator of $\mathbf{D}$. Hence, $\mathbf{P}^{-1}\mathbf{D}_2\mathbf{P}^{-1'}$ of (74) is an unbiased estimator of $\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\mathbf{D}\mathbf{P}^{-1'}$.

## 10. OTHER WAYS TO ESTIMATE VARIANCE COMPONENTS

This section will give a sketch of three other methods of variance components estimation. These are maximum likelihood estimation, Bayes estimation, and the MINQUE method. Most of the discussion will be confined to the one-way classification. We will present a few results for the two-way classification without proof. For a more extensive treatment of these topics, the reader may consult Searle, Casella, and Mc Culloch (1992) and the references therein. An overview of these topics is available in Searle (1995).

### a. Maximum Likelihood Estimation

Estimating parameters of a fixed-effects model by the method of maximum likelihood, under normality assumptions, frequently leads to the same estimators as do least squares and best linear unbiased estimation. However, for variance component estimators, it does not lead to the analysis of variance estimators. The analysis of variance estimators cannot be maximum likelihood estimators because as we have already seen, they can be negative. We obtain maximum likelihood estimators by maximizing the likelihood over a parameter space that is non-negative as far as the variance components are concerned. Therefore, maximum likelihood estimators must be non-negative. The derivation of maximum likelihood estimators is not as straight forward for variance components as it is for the parameters of a fixed-effects model. Indeed, with unbalanced data, explicit estimators cannot be obtained. We now discuss some of the available results for balanced data in the one-way classification.

We first consider the un-restricted maximum likelihood estimator and then consider the restricted maximum likelihood estimator for the one-way classification. We then give the maximum likelihood estimators for the two-way classification without derivation.

**(*i*) *The Unrestricted Maximum Likelihood Estimator.*** The likelihood of the sample of observations in the one-way classification model is

$$L = (2\pi)^{-\frac{1}{2}an}|\mathbf{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu\mathbf{1})'\mathbf{V}^{-1}(\mathbf{y} - \mu\mathbf{1}). \tag{75}$$

The matrix $\mathbf{V}$ of (45) may be rewritten as

$$\mathbf{V} = \sum_{i=1}^{a} {}^{+}(\sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J}),$$

where the order of $\mathbf{I}$ and $\mathbf{J}$ is $n$. Then the determinant of $\mathbf{V}$ is

$$|\mathbf{V}| = \prod_{i=1}^{a} |(\sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J})| = [\sigma_e^{2(n-1)}(\sigma_e^2 + n\sigma_\alpha^2)]^a$$

and the inverse of $\mathbf{V}$ is

$$\mathbf{V}^{-1} = \sum_{i=1}^{a} {}^{+}\left[\frac{1}{\sigma_e^2}\mathbf{I} - \frac{\sigma_\alpha^2}{\sigma_e^2(\sigma_e^2 + \sigma_\alpha^2)}\mathbf{J}\right].$$

Substituting for $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ into (75), after some simplification yields

$$L = \frac{\exp -\frac{1}{2}\left[\dfrac{\text{SSE}}{\sigma_e^2} + \dfrac{\text{SSA}}{\sigma_e^2 + n\sigma_\alpha^2} + \dfrac{an(\bar{y}_{\cdot\cdot} - \mu)^2}{\sigma_e^2 + n\sigma_\alpha^2}\right]}{(2\pi)^{\frac{1}{2}an}(\sigma_e^2)^{\frac{1}{2}a(n-1)}(\sigma_e^2 + n\sigma_\alpha^2)^{\frac{1}{2}a}}. \tag{76}$$

Equating to zero, the derivatives of $\log L$ with respect to $\mu$, $\sigma_\alpha^2$, and $\sigma_e^2$ and denoting the solutions by $\tilde{\mu}$, $\tilde{\sigma}_\alpha^2$, and $\tilde{\sigma}_e^2$ gives $\tilde{\mu} = \bar{y}_{\cdot\cdot}$ and

$$a(\tilde{\sigma}_e^2 + n\tilde{\sigma}_\alpha^2) = \text{SSA} \quad \text{and} \quad a(n-1)\tilde{\sigma}_e^2 = \text{SSE}. \tag{77}$$

The solutions to (77) are

$$\tilde{\sigma}_e^2 = \frac{\text{SSE}}{a(n-1)} = \text{MSE} \tag{78a}$$

and

$$\tilde{\sigma}_\alpha^2 = \frac{\text{SSA}/a - \tilde{\sigma}_e^2}{n} = \frac{(1 - 1/a)\,\text{MSA-MSE}}{n} \tag{78b}$$

When maximizing $L$ in (76), we did not restrict the parameters $\tilde{\sigma}_\alpha^2$ and $\tilde{\sigma}_e^2$ to positive values. The solutions to (77) given in (78a) and (78b) are not maximum likelihood estimators. Herbach (1959) and Searle, Casella, and Mc Culloch (1992, pp. 81–84), show that when $\tilde{\sigma}_\alpha^2$ is negative or equivalently (1 – 1/a)MSA < MSE, the maximum likelihood estimator of $\sigma_\alpha^2$ is zero and that of $\sigma_e^2$ is SST/*an*.

**(*ii*) *Restricted Maximum Likelihood Estimator*** We now consider an adaption of maximum likelihood estimation, which maximizes that portion of the likelihood confined to sufficient statistics that are location invariant (see pp. 296–300 of Casella and Berger (2002)). For the one-way classification, this means maximizing that portion of the likelihood that does not involve $\mu$.

Thompson (1962) and Patterson and Thompson (1971) suggest restricted maximum likelihood estimation. Anderson and Bancroft (1952, p. 320) and Russell and Bradley (1958) consider similar estimation procedures. Thus, we maximize

$$L = \frac{\exp -\frac{1}{2}\left[\frac{\text{SSE}}{\sigma_e^2} + \frac{\text{SSA}}{\lambda}\right]}{(2\pi)^{\frac{1}{2}(an-1)}(\sigma_e^2)^{\frac{1}{2}a(n-1)}(\lambda)^{\frac{1}{2}(a-1)}(an)^{\frac{1}{2}}}, \tag{79}$$

where $\lambda = \sigma_e^2 + n\sigma_\alpha^2$. Finding the partial derivatives of $\log L$ in (79) with respect to $\sigma_e^2$ and $\lambda$, equating them to zero and solving the resulting equations, we find that restricted maximum likelihood solutions are

$$\tilde{\sigma}_{e,R}^2 = \frac{\text{SSE}}{n-1} = \text{MSE} \quad \text{and} \quad \tilde{\sigma}_{\alpha,R}^2 = \frac{1}{n}(\text{MSA} - \text{MSE}). \tag{80a}$$

From considerations similar to those in finding the unrestricted maximum likelihood estimator, we have that the restricted maximum likelihood estimator is given by (80a) when $\tilde{\sigma}_{\alpha,R}^2 > 0$ and when $\tilde{\sigma}_{\alpha,R}^2 \leq 0$,

$$\tilde{\sigma}_{e,R}^2 = \frac{\text{SST}}{an-1} \quad \text{and} \quad \tilde{\sigma}_{\alpha,R}^2 = 0. \tag{80b}$$

**Example 14 Illustration of Numerical Values of the Maximum Likelihood Estimator** For the computer output in Example 2, we have $a = 6$, $n = 6$, MSA $= 1.960$, and MSE $= 0.313$. Then $\tilde{\sigma}_e^2 = 0.313$ and $\tilde{\sigma}_\alpha^2 = \frac{1}{6}\left[\left(1 - \frac{1}{6}\right)1.960 - 0.313\right] = 0.220$ is the maximum likelihood estimate. The restricted maximum likelihood estimator is the same as the analysis of variance estimator that was obtained in Example 2.

The results of this section are summarized in Table 9.15 below. □

**(*iii*) *The Maximum Likelihood Estimator in the Two-Way Classification.*** For a two-way model where both factors are random, the maximum likelihood estimators cannot be found in closed form. They can be found in closed form for the mixed model however. These closed forms are given in Tables 9.16–9.18 where A is the fixed factor and B is the random factor.

Searle, Casella, and Mc Culloch (1992, p. 253) point out that solutions to the restricted maximum likelihood estimators in balanced data for all mixed models are the same as that of the analysis of variance estimator. Anderson (1978, pp 97–104)

**TABLE 9.15 Estimators of Variance Components in the One-Way Classification, Random Model, With Balanced Data**

| Methods of Estimation | Conditions | Estimators of $\sigma_\alpha^2$ | Estimators of $\sigma_e^2$ |
|---|---|---|---|
| Analysis of variance | None | $\frac{(\text{MSA} - \text{MSE})}{n}$ | MSE |
| Maximum likelihood (Herbach, 1959) | $\frac{a-1}{a}\text{MSA} \geq \text{MSE}$ | $\frac{(((a-1)/a)\,\text{MSA} - \text{MSE})}{n}$ | MSE |
| | $\frac{a-1}{a}\text{MSA} < \text{MSE}$ | 0 | $\frac{\text{SST}}{an}$ |
| Restricted maximum likelihood (Thompson, 1962) | MSA $\geq$ MSE | $\frac{(\text{MSA} - \text{MSE})}{n}$ | MSE |
| | MSA < MSE | 0 | $\frac{\text{SST}}{an-1}$ |

gives a detailed proof of this result. For discussion of this result and some specific examples, see Patterson and Thompson (1971), Corbeil and Searle (1976), Searle (1976), and Harville (1977).

### b. The MINQUE

We give a brief summary of the MINQUE method of estimating variance components. It is particularly useful for unbalanced data because its variance is less than that of the analysis of variance estimator. Also for certain modifications of the MINQUE estimator, we can avoid negative variance components. For the most part, we follow P. S. R. S. Rao (1997).

**TABLE 9.16 Maximum Likelihood (ML) Estimators of $\sigma_\alpha^2$, $\sigma_\beta^2$, and $\sigma_\gamma^2$ in a Two-Way Nested Classification Mixed Model**

| Conditions satisfied by the ML soluations | $\tilde{\sigma}_\alpha^2$ | MLE $\tilde{\sigma}_\beta^2$ | $\tilde{\sigma}_e^2$ |
|---|---|---|---|
| $(a-1)\,\text{MSA} \geq a\,\text{MSB}$: A, MSB: A $\geq$ MSB | $\frac{\text{SSA} - a\,\text{MSB: A}}{abn}$ | $\frac{\text{MSB: A} - \text{MSE}}{n}$ | MSE |
| $(a-1)\,\text{MSA} \geq a\,\text{MSB}$: A, MSB: A < MSB | $\frac{\text{SSA} - a\tilde{\sigma}_e^2}{abn}$ | 0 | $\frac{\text{SSE} + \text{SSB: A}}{a(bn-1)}$ |
| $(a-1)\,\text{MSA} < a\,\text{MSB}$: A, MSB: A $\geq$ MSB | 0 | $\left(\frac{\text{SSA} + \text{SSB: A} - ab\text{MSE}}{abn}\right)$ | MSE |
| $(a-1)\,\text{MSA} < a\,\text{MSB}$: A, MSB: A < MSB | 0 | 0 | $\frac{\text{SST}_\text{m}}{abn}$ |

**TABLE 9.17 Maximum Likelihood Estimators of $\sigma_\alpha^2, \sigma_\beta^2$, and $\sigma_\gamma^2$ in the Two-Way Crossed Classification Mixed Model**

| Conditions Satisfied by the ML Solutions | $\tilde{\sigma}_\beta^2$ | MLE $\tilde{\sigma}_\gamma^2$ | $\tilde{\sigma}_e^2$ |
|---|---|---|---|
| $(b-1)\,\text{MSB} \geq \text{MSAB}, (b-1)\,\text{MSAB} \geq \text{MSE}$ | $\dfrac{(b-1)(\text{MSB}-\text{MSAB})}{abn}$ | $\dfrac{(b-1)(\text{MSAB}-\text{MSE})}{bn}$ | MSE |
| $(b-1)\,\text{MSB} \geq \text{MSAB}, (b-1)\,\text{MSAB} < \text{MSE}$ | $\dfrac{1}{an}\left(\dfrac{\text{SSB}}{b}+\dfrac{\text{SSE}+\text{SSAB}}{abn-a-b+1}\right)$ | 0 | $\dfrac{\text{SSB}}{b}+\dfrac{\text{SSE}+\text{SSAB}}{abn-a-b+1}$ |
| $(b-1)\,\text{MSB} < \text{MSAB}, (b-1)\,\text{MSAB} \geq \text{MSE}$ | 0 | $\dfrac{\sum_{i=1}^{a}\sum_{j=1}^{b} n(\bar{y}_{ij\cdot}-\bar{y}_{\cdots})^2}{ab-1}$ | MSE |
| $(b-1)\,\text{MSB} < \text{MSAB}, (b-1)\,\text{MSAB} < \text{MSE}$ | 0 | 0 | $\dfrac{\text{SST}_\text{m}}{abn}$ |

**TABLE 9.18 Maximum Likelihood Estimators of $\sigma_\beta^2$ and $\sigma_e^2$ in a Two-Way Crossed Classification, Mixed Model, $\alpha$ Effects Fixed**

| Conditions Satisfied by the ML solutions | MLE $\tilde{\sigma}_\beta^2$ | $\tilde{\sigma}_e^2$ |
|---|---|---|
| $\frac{(b-1)}{b}\text{MSB} \geq \left[1 - \frac{(a-1)}{b(an-1)}\right]\text{MSE}$ | $\frac{\frac{(b-1)}{b}\text{MSB} - \left[1 - \frac{a-1}{b(an-1)}\text{MSE}\right]}{an}$ | $\left[1 - \frac{a-1}{b(an-1)}\text{MSE}\right]$ |
| $\frac{(b-1)}{b}\text{MSB} < \left[1 - \frac{(a-1)}{b(an-1)}\right]\text{MSE}$ | 0 | $\frac{\text{SST}_\text{m}}{abn}$ |

**(*i*) *The Basic Principle.*** The letters in MINQUE stand for minimum, invariant, norm, quadratic unbiased estimation, respectively. In the course of this development, we shall explain how this comes into play.

Consider a linear model

$$\mathbf{Y} = \mathbf{Xb} + \varepsilon \tag{81}$$

with $\mathbf{Y}$ an $n \times 1$ vector of observations, $\mathbf{X}$ a known $n \times s$ matrix, and $\beta$ an $s \times 1$ vector of fixed parameters. We can express the $n \times 1$ vector $\mathbf{e}$ as

$$\varepsilon = \mathbf{U}_1\xi_1 + \mathbf{U}_2\xi_2 + \cdots + \mathbf{U}_p\xi_p = \mathbf{U}\xi. \tag{82}$$

The $n \times n_i$ matrices $\mathbf{U}_i, i = 1, 2, \ldots, p$ are known and the $n_i \times 1$ vectors $\xi_i$ represent the random effects and residuals $n = \sum_{i=1}^{p} n_i$.. Observe that $\mathbf{U} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2 & \cdots & \mathbf{U}_p\end{bmatrix}, \xi' = \begin{bmatrix}\xi_1', \xi_2', \cdots, \xi_p\end{bmatrix}$ and $\mathbf{e} = \mathbf{U}\xi$.

We assume that $\xi_i$ follows a normal distribution with $E(\xi_i) = 0, E(\xi_i\xi_i') = \sigma_i^2\mathbf{I}_i$, and $E(\xi_i\xi_j') = 0$ for $i \neq j$. The variance–covariance matrix or dispersion matrix of $\mathbf{Y}$ is

$$\Sigma = E(\varepsilon\varepsilon') = E(\mathbf{U}\xi)(\mathbf{U}\xi)' = \sum_{i=1}^{p} \sigma_i^2\mathbf{V}_i, \tag{83a}$$

where $\mathbf{V}_i = \mathbf{U}_i\mathbf{U}'_i$. When $\xi_i$, $i = 1,2, p$ has a normal distribution, $\mathbf{Y}$ follows a multinormal distribution with mean $\mathbf{X}\beta$ and dispersion $\Sigma$.

**Example 15 What this Model Looks Like for the Balanced One-Way Analysis of Variance** For the balanced one-way analysis of variance $\mathbf{Y}$ is an $na \times 1$ vector of observations, $\mathbf{X} = \mathbf{1}_{na}$, $\mathbf{U}_1$ is the $na \times a$ matrix

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{1}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_n \end{bmatrix}$$

and $\boldsymbol{\xi}_1 = \alpha$. Now $\mathbf{U}_2$ is the $na \times na$ matrix $\mathrm{I}_{na}$ and $\boldsymbol{\xi}_2 = \mathbf{e}$. Then,

$$\mathbf{V}_1 = \begin{bmatrix} \mathbf{J}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_n & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \mathbf{J}_n \end{bmatrix}$$

and $\mathbf{V}_2 = \mathrm{I}_{na.}$ We then have that

$$\boldsymbol{\xi} = \mathbf{U}_1\alpha + \mathbf{U}_2\mathbf{e}$$

and

$$\mathbf{V} = \sigma_\alpha^2\mathbf{V}_1 + \sigma_e^2\mathbf{V}_2.$$

Consider a linear combination of the variance components of the model in (81) in the form

$$\mathbf{l}'\sigma = l_1\sigma_1^2 + l_2\sigma_2^2 + \cdots + l_p\sigma_p^2 \tag{83b}$$

where $\mathbf{l}$ is a $p \times 1$ column vector with elements $l_i$, $1 \le i \le p$ and $\sigma$ is a $p \times 1$ column vector with elements $\sigma_i^2, 1 \le i \le p$.

To estimate this linear combination we consider we use a quadratic form of the observations that is invariant to $\boldsymbol{\beta}$ . For any $s \times 1$ vector $\boldsymbol{\beta}_0$ the quadratic form

$$(\mathbf{Y} - \mathbf{X}\beta_0)'\mathbf{A}(\mathbf{Y} - \mathbf{X}\beta_0) = \mathbf{Y}'\mathbf{A}\mathbf{Y} \tag{84}$$

provided that $\mathbf{AX} = \mathbf{0}$. Note that $\mathbf{AX} = \mathbf{0}$ implies $\mathbf{X}'\mathbf{AX} = \mathbf{0}$. From (84), using translation invariance, we have that

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \boldsymbol{\varepsilon}'\mathbf{A}\boldsymbol{\varepsilon} = \boldsymbol{\xi}'\mathbf{U}'\mathbf{A}\mathbf{U}\boldsymbol{\xi} \tag{85}$$

and from (83a),

$$E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = Etr(\mathbf{A}\mathbf{U}\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{U}') = \sum_{i=1}^{p} \sigma_i^2 \mathrm{tr}\mathbf{A}\mathbf{V}_i. \tag{86}$$

Hence, $\mathbf{l}'\hat{\sigma} = \mathbf{Y}'\mathbf{A}\mathbf{Y}$ will be unbiased for $\mathbf{l}'\sigma$ if $\mathrm{tr}\mathbf{A}\mathbf{V}_i = \mathbf{l}_i$ for $i = 1, 2, \ldots, p$.

To formulate the MINQUE principle, Rao (1971a) starts with the estimator

$$\left(\frac{l_1}{n_1}\right)\boldsymbol{\xi}_1'\boldsymbol{\xi}_1 + \left(\frac{l_2}{n_2}\right)\boldsymbol{\xi}_2'\boldsymbol{\xi}_2 + \cdots + \left(\frac{l_p}{n_p}\right)\boldsymbol{\xi}_p'\boldsymbol{\xi}_p = \boldsymbol{\xi}'\boldsymbol{\Delta}\boldsymbol{\xi}. \tag{87}$$

The matrix $\boldsymbol{\Delta}$ is block diagonal with diagonal blocks $(l_i/n_i)\mathbf{I}_i$, where $\mathbf{I}_i$ is the identity matrix of dimension $n_i$.

From (85) and (87), we obtain the difference

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} - \boldsymbol{\xi}'\boldsymbol{\Delta}\boldsymbol{\xi} = \boldsymbol{\xi}'(\mathbf{U}'\mathbf{A}\mathbf{U} - \boldsymbol{\Delta})\boldsymbol{\xi} \tag{88}$$

We reduce this distance by minimizing the square of the Euclidean norm

$$\begin{aligned} \left\|\mathbf{U}'\mathbf{A}\mathbf{U} - \boldsymbol{\Delta}\right\|^2 &= \mathrm{tr}(\mathbf{U}'\mathbf{A}\mathbf{U} - \boldsymbol{\Delta})^2 \\ &= \mathrm{tr}(\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V} + \boldsymbol{\Delta}^2 - 2\mathbf{A}\mathbf{U}\boldsymbol{\Delta}\mathbf{U}') \\ &= \mathrm{tr}\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V} - \mathrm{tr}\boldsymbol{\Delta}^2. \end{aligned} \tag{89}$$

Note that $\mathrm{tr}\boldsymbol{\Delta}^2 = \sum_{i=1}^{p} (l_i^2/n_i)$. The last expression in (89) is obtained by the unbiasedness condition.

From the development above, finding the matrix $\mathbf{A}$ of MINQUE of $\mathbf{l}'\sigma = \mathbf{Y}'\mathbf{A}\mathbf{Y}$ consists of minimizing $\mathrm{tr}\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V}$ with the variance and unbiasedness conditions

$$(i)\ \mathbf{A}\mathbf{X} = \mathbf{0} \quad \text{and} \quad (ii)\ \mathrm{tr}\mathbf{A}\mathbf{V}_i = l_i. \tag{90}$$

**(*ii*) *The MINQUE Solution*** Observe that

$$\mathbf{A}\mathbf{X} = \mathbf{0} \Rightarrow \mathbf{B}\mathbf{X}_0 = \mathbf{0} \tag{91}$$

where $\mathbf{B} = \mathbf{V}^{1/2}\mathbf{A}\mathbf{V}^{1/2}$ and $\mathbf{X}_0 = \mathbf{V}^{-1/2}\mathbf{X}$. The condition for unbiasedness now becomes

$$\mathrm{tr}\mathbf{A}\mathbf{V}_i = \mathrm{tr}\mathbf{B}\mathbf{V}^{-1/2}\mathbf{V}_i\mathbf{V}^{-1/2} = l_i. \tag{92}$$

Thus, minimizing $\mathrm{tr}\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{V}$ with the two conditions in (90) is the same as minimizing $\mathrm{tr}\mathbf{B}^2$ with the conditions in (91) and (92). Now,

$$\mathbf{B}\mathbf{X}_0 = \mathbf{0} \Rightarrow \mathbf{B} = \mathbf{Q}_0\mathbf{B}\mathbf{Q}_0 \tag{93}$$

where $\mathbf{Q}_0 = \mathbf{I} - \mathbf{X}_0(\mathbf{X}'_0\mathbf{X}_o)^{-}\mathbf{X}'_0$. As a result,

$$\mathrm{tr}\mathbf{B}\mathbf{V}^{-1/2}\mathbf{V}_i\mathbf{V}^{-1/2} = \mathrm{tr}\mathbf{B}\mathbf{Q}_0\mathbf{V}^{-1/2}\mathbf{V}_i\mathbf{V}^{-1/2}\mathbf{Q}_0 = l_i. \tag{94}$$

The solution for minimizing $\mathrm{tr}\mathbf{B}^2$ with the constraints in (93) and (94) is

$$\mathbf{B} = \sum_{i=1}^{p} \lambda_i \mathbf{Q}_0\mathbf{V}^{-1/2}\mathbf{V}_i\mathbf{V}^{-1/2}\mathbf{Q}_0. \tag{95}$$

Thus, we have that

$$\mathbf{A} = \sum_{i=1}^{p} \lambda_i \mathbf{R}_0\mathbf{V}_i\mathbf{R}_0, \tag{96}$$

where $\mathbf{R}_0 = \mathbf{V}^{-1/2}\mathbf{Q}_0\mathbf{V}^{-1/2} = \mathbf{V}^{-1}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]$. We obtain the coefficients $\lambda_i$ from the unbiasedness condition in (90), that is from

$$\sum_{i=1}^{p} \lambda_i \mathrm{tr}\mathbf{R}_0\mathbf{V}_i\mathbf{R}_0\mathbf{V}_j = l_j, j = 1, 2, \ldots, p. \tag{97}$$

(See Section 1f.3, of C. R. Rao (1973, pp. 65–66).)

Equations (97) may be expressed as $\mathbf{F}_0\lambda = l$ where $\lambda$ and $\mathbf{l}$ are the vectors of $\lambda_i$ and $l_i$, respectively.

From (96), we see that the MINQUE of $\mathbf{l}'\sigma$ is

$$\mathbf{l}'\hat{\sigma} = \sum_{i=1}^{p} \lambda_i e_0'\mathbf{V}_i e_0 = \sum_{i=1}^{p} \lambda_i g_{0i} = \lambda'\mathbf{g}_0, \tag{98}$$

where $\mathbf{e}_0 = \mathbf{R}_0\mathbf{Y}, g_{0i} = \mathbf{e}'_0\mathbf{V}_i\mathbf{e}_0$ and $\mathbf{g}_0$ is the vector of the elements $g_{0i}$ for $i = 1, 2, \ldots, p$. Since $\mathbf{l}' = \lambda'\mathbf{F}_0$ it follows from (98), that

$$\mathbf{F}_0\hat{\sigma} = \mathbf{g}_0. \tag{99}$$

Thus, we obtain the MINQUE of the individual variance components and $\mathbf{l}'\sigma$ from the equations $\hat{\sigma} = \mathbf{F}_0^{-1}\mathbf{g}_0$ and $\mathbf{l}'\hat{\sigma} = \mathbf{l}'\mathbf{F}_0^{-1}\mathbf{g}_0$.

This procedure for estimating the variance components is equivalent to equating $\mathbf{e}'\mathbf{V}_i\mathbf{e}$ to its expectation and solving for $\sigma_i^2$. From (98) and (99), we see that the MINQUE is a linear combination of the quadratic forms of the residuals $\mathbf{Q}_0\mathbf{Y}$ obtained by regressing $\mathbf{V}^{-1/2}\mathbf{Y}$ on $\mathbf{V}^{-1/2}\mathbf{X}$.

Alternative derivations of the MINQUE are available in P. S. R. S. Rao (1997), C. R. Rao (1973), C. R. Rao (1984), Mitra (1971), and Brown (1977).

**(*iii*) *A priori Values and the MIVQUE*** Suppose a priori values are available for $\sigma_i^2, i = 1, 2, \ldots, p$ for the model in (82). Denote these values by $\gamma_i^2$. For the model (82), we have,

$$\varepsilon = U_{1*}\eta_1 + U_{2*}\eta_2 + \cdots + U_{p*}\eta_p \tag{100}$$

with $U_{i*} = U_i\gamma_i, U_* = (U_{1*}, U_{2*}, \ldots, U_{p*}), \eta_i = (1/\gamma_i)\xi_i$ and $\eta' = (\eta_1', \eta_2', \ldots, \eta_p')$. Observe that $\mathbf{U}_* = \mathbf{U}\Lambda^{1/2}$ and $\eta = \Lambda^{-1/2}\xi$ where $\Lambda^{-1/2}$ is a diagonal matrix with elements $\gamma_i\mathbf{I}_i$.

Using (100) and the invariance condition $\mathbf{AX} = \mathbf{0}$, we have

$$\mathbf{Y}'\mathbf{AY} = \xi'\mathbf{U}'\mathbf{AU}\xi = \eta'(\Lambda^{1/2}\mathbf{U}'\mathbf{AU}\Lambda^{1/2})\eta. \tag{101}$$

We may express the estimator $\sum_{i=1}^{p} (l_i/n_i)\boldsymbol{\xi}_i\boldsymbol{\xi}_i'$ in (87) as

$$\boldsymbol{\xi}'\boldsymbol{\Delta}\boldsymbol{\xi} = \boldsymbol{\eta}'(\boldsymbol{\Lambda}^{1/2}\mathbf{U}'\mathbf{A}\mathbf{U}\boldsymbol{\Lambda}^{1/2})\boldsymbol{\eta}. \tag{102}$$

To minimize the difference between (101) and (102), consider

$$\begin{aligned} &\mathrm{tr}[\boldsymbol{\Lambda}^{1/2}(\mathbf{U}'\mathbf{A}\mathbf{U} - \boldsymbol{\Delta})\boldsymbol{\Lambda}^{1/2}] \\ &\quad = \mathrm{tr}\mathbf{A}(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}')\mathbf{A}(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}') + \mathrm{tr}\boldsymbol{\Delta}\boldsymbol{\Lambda}\boldsymbol{\Delta}\boldsymbol{\Lambda} - 2\mathrm{tr}\mathbf{A}\mathbf{U}(\boldsymbol{\Lambda}\boldsymbol{\Delta}\boldsymbol{\Lambda})\mathbf{U}' \\ &\quad = \mathrm{tr}\mathbf{A}\mathbf{V}_*\mathbf{A}\mathbf{V}_* - \sum_{i=1}^{p}\left(\frac{l_i}{n_i}\right)\gamma_i^2, \end{aligned} \tag{103}$$

where $\mathbf{V}_* = \mathbf{U}_*\mathbf{U}'_*$. We obtain the last expression through the unbiasedness condition.

We obtain the MINQUE by minimizing $\mathrm{tr}\mathbf{A}\mathbf{V}_*\mathbf{A}\mathbf{V}_*$ together with the constraints in (90). The resulting solution is

$$\mathbf{A} = \sum_{i=1}^{p} \lambda_i \mathbf{R}\mathbf{V}_i\mathbf{R}, \tag{104}$$

where $\mathbf{R} = \mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}$. We obtain the coefficients $\lambda_i$ from

$$\sum_{i=1}^{p} \lambda_i \mathrm{tr}\mathbf{R}\mathbf{V}_i\mathbf{R}\mathbf{V}_j = l_j, j = 1, 2, \ldots, p. \tag{105}$$

We have as before

$$\mathbf{l}'\hat{\sigma} = \sum_{i=1}^{p} \lambda_i \mathbf{e}'\mathbf{V}_i\mathbf{e} = \sum_{i=1}^{p} \lambda_i g_i, \tag{106}$$

where $\mathbf{e} = \mathbf{R}\mathbf{Y}$ and $g_i = \mathbf{e}'\mathbf{V}_i\mathbf{e}$. As before, we obtain the MINQUEs from

$$\mathbf{F}\hat{\sigma} = \mathbf{g} \tag{107}$$

where the elements of $\mathbf{F}$ are $f_{ij} = \mathrm{tr}\mathbf{R}\mathbf{V}_i\mathbf{R}\mathbf{V}_j$.

The MIVQUE is an estimator obtained by minimizing the variance instead of the norm subject to the same unbiasedness and invariance conditions as for the MINQUE.

Methods of finding the MIVQUE from (99) and (107) are available in Lou and Senturia (1977), P. S. R. S. Rao, Kaplan, and Cochran (1983), Kaplan (1983), Giesbrecht (1983), Kleffe and Seifert (1984), and others.

**(*iv*) *Some Properties of the MINQUE*** The MINQUE has some other desirable properties besides invariance and unbiasedness. These include

1. When all the $\boldsymbol{n}_i$ elements of $\boldsymbol{\xi}_i$ have the same variance $\sigma_i^2$ and the same fourth moment $\boldsymbol{\mu}_{4i}$ with the invariance condition $\mathbf{AX} = 0$**,**

$$V(\mathbf{Y}'\mathbf{AY}) = 2\text{tr}\mathbf{A}\Sigma\mathbf{A}\Sigma + \sum_{i=1}^{p} (\boldsymbol{\mu}_{4i} - 3\sigma_i^4)\text{tr}\mathbf{AV}_i\mathbf{AV}_i. \tag{108}$$

   If $\boldsymbol{\xi}_i$ is normally distributed, the second term of (108) vanishes. Then the MINQUE obtained in (ii) is the same as the estimator obtained by minimizing (108) with $\mathbf{V}_* = \sum_{i=1}^{p} \gamma_i^2\mathbf{V}_i$ in place of $\Sigma$. As a result, the MINQUE is the same as the MIVQUE with a priori values.
2. The MINQUE in (98) is a linear combination of the quadratic forms of the residuals $\mathbf{Q}_*\mathbf{Y}$. We can find these residuals by regressing $\mathbf{V}_*^{-1/2}\mathbf{Y}$ on $\mathbf{V}_*^{-1/2}\mathbf{X}$.
3. If $\boldsymbol{\gamma}_i^2$ are close to $\sigma_i^2$ for $i = 1, 2, \ldots, p$ the restricted maximum likelihood estimators and the MINQUE estimators for $\sigma_i^2$ with constraints for non-negativeness would be expected to be the same.
4. When prior observations $\gamma_i^2$ are not available we can obtain the MINQUEs for $\sigma_i^2$ by iterative procedures (See Section 6.8 P. S. R. S. Rao (1997)).

**Example 16 The MIVQUE of a Common Means Model** Consider the model

$$y_{ij} = \mu + \varepsilon_{ij}, i = 1, 2, \ldots, k \quad \text{and} \quad j = 1, 2, \ldots, n_i \tag{109}$$

with $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma_i^2$. This model represents observations from samples of sizes $n_i$ from $k$ populations with a common mean $\mu$ but different variances $\sigma_i^2$. We assume that the residuals $\varepsilon_{ij}$ within a group and among groups are uncorrelated.

We find that

$$\gamma_i^4 e'\mathbf{V}_i e = (n_i - 1)s_i^2 + n_i(\bar{y}_i - \bar{y}_W)^2, \tag{110}$$

where $\bar{y}_i$ and $s_i^2$ are the sample mean and variance of the $i$th group. Furthermore, $W_i = n_i/\gamma_i$, $W = \sum_{i=1}^{k} W_i$, and $\bar{y}_W = \sum_{i=1}^{k} W_i\bar{y}_i/W$. Finding the expectation of (110), we have that

$$\gamma_i^4 E(e'\mathbf{V}_i e) = \left(n_i - 2\frac{W_i}{W}\right)\sigma_i^2 + \frac{n_i}{W^2}\sum_{i=1}^{k}\frac{W_i^2\sigma_i^2}{n_i}. \tag{111}$$

To find the MIVQUE, equate the right-hand sides of (110) and (111) and solve for $\sigma_i^2$. Numerical methods would have to be used. However, we will obtain an estimate of $\sigma_i^2$ by making a simplifying assumption.

From equations (110) and (111), we see that the MIVQUEs depend only on the relative values of the a priori values. If there is no information available about the variances all of the $\gamma_i^2$ may be considered equal to unity. Then equating the right-hand sides of (110) and (111) for each $i$ and solving the resulting system of equations we find that

$$\hat{\sigma}_i^2 = \frac{1}{n-2}\left[\frac{n}{n_i}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}{n-1}\right]. \tag{112}$$

Suppose we have only one sample and $n_i = n$. We can drop the subscript $i$ and obtain the well-known unbiased estimate of the variance

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})^2. \tag{113}$$

For a single random sample of $n$ observations, the MIVQUE is the usual unbiased estimate of the variance.

**(*v*) *Non-negative Estimators of Variance Components*** P. S. R. S. Rao and Chaubey (1978) show that we can find non-negative estimates of variance components (MINQE) by ignoring the unbiasedness condition. Using the notation of the previous sub-section taking the derivative of the Euclidean norm with respect to $\mathbf{A}$ and setting it to zero we get

$$\mathbf{V}_*\mathbf{A}\mathbf{V}_* = \mathbf{U}_*\mathbf{\Lambda}^{1/2}\mathbf{\Delta}\mathbf{\Lambda}^{1/2}\mathbf{U}'_*. \tag{114}$$

(See C. R. Rao (1973), Exercise 13.2, p. 72.)

Using the invariance condition $\mathbf{AX} = \mathbf{0}$, we have, using notation already defined, that

$$\mathbf{V}_*^{1/2}\mathbf{A}\mathbf{V}_*^{1/2} = \mathbf{Q}_*\mathbf{V}_*^{1/2}\mathbf{A}\mathbf{V}_*^{1/2}\mathbf{Q}_* \tag{115}$$

Substitution of (115) into (114) gives

$$\mathbf{A} = \mathbf{R}\mathbf{U}_*\mathbf{\Lambda}^{1/2}\mathbf{\Delta}\mathbf{\Lambda}^{1/2}\mathbf{U}'_*\mathbf{R} = \sum_{i=1}^{p} l_i\left(\frac{\gamma_i^4}{n_i}\right)\mathbf{R}\mathbf{V}_i\mathbf{R}. \tag{116}$$

Hence the MINQE of $\sigma_i^2$ is

$$\hat{\sigma}_i^2 = \left(\frac{\gamma_i^4}{n_i}\right) e'\mathbf{V}_i e = \left(\frac{\gamma_i^4}{n_i}\right) g_{i.} \tag{117}$$

The expression in (117) is non-negative. □

**Example 17 The MINQE for a Model with Common Mean Discussed in Example 16** For the model in (109), the MINQE for $\sigma_i^2$ takes the form

$$\hat{\sigma}_i^2 = \frac{[(n_i - 1)s_i^2 + n_i(\bar{y}_i - \bar{y}_W)^2]}{n_i} \tag{118}$$

$W_i = n_i/\gamma_i^2$ and $\bar{y}_W = \sum_{i=1}^k W_i\bar{y}_i / \sum_{i=1}^k W_i$. If $\sigma_i^2 = \sigma^2$ the MINQE for $\sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2}{n}, \tag{119}$$

the average of the squared residuals. □

### c. Bayes Estimation

Bayesian statistics makes use of both the data collected as a result of a statistical experiment and prior information about the parameter we wish to estimate. In order to be able to do this, we combine the prior information with the information obtained in sampling using Bayes Theorem to calculate a posterior distribution. We will illustrate this, first by a simple example and then by estimating the variance components for the one-way balanced model. The discussion here is adapted from that of Searle, Casella and Mc Culloch (1992).

(*i*) ***Bayes Theorem and the Calculation of a Posterior Distribution*** Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random sample from a population. Let $\pi(\theta)$ be a prior distribution of the parameter $\theta$. Then Bayes Theorem states that the posterior distribution is given by

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x}|\theta)\pi(\theta)d\theta} \tag{120}$$

The joint distribution of the components of $\mathbf{x}$ is called the likelihood functions and $\Theta$ represents all of the possible values of the parameter $\theta$.

We will illustrate the calculation of a posterior distribution by means of a simple example.

**Example 18 Posterior Distribution of a Variance** Assume that $\mathbf{x} \sim N(\mu\mathbf{1}, \sigma^2\mathbf{I})$. The usual unbiased estimator of $\sigma^2$ is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \quad \text{with} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{121}$$

We then have,

$$f(s^2|\sigma^2) = \frac{(m/\sigma^2)^{\frac{1}{2}m} s^{2\left(\frac{1}{2}m-1\right)} e^{-\frac{1}{2}ms^2/\sigma^2}}{\mathbf{\Gamma}\left(\frac{1}{2}m\right) 2^{\frac{1}{2}m}}, \tag{122}$$

where $m = n - 1$.

For the prior distribution, we will use the inverted gamma distribution. We can obtain it by finding the distribution of the reciprocal of a gamma random variable. Its general form is

$$f(x) = \frac{x^{-(a+1)} e^{-1/bx}}{\mathbf{\Gamma}(a) b^a}. \tag{123}$$

We shall use this with $a = 2$ and $b = 1$ as the prior for $\sigma^2$. Because the resulting distribution will have infinite variance, so rather vague prior information will be imparted.

Using (120), we need to calculate

$$\pi(\sigma^2, s^2) = \frac{f(s|\sigma^2)\pi(\sigma^2)}{\int_0^\infty f(s|\sigma^2)\pi(\sigma^2)d\sigma^2}. \tag{124}$$

The numerator is

$$f(s^2|\sigma^2)\pi(\sigma^2) = \frac{m^{\frac{1}{2}m}}{\mathbf{\Gamma}\left(\frac{1}{2}m\right) 2^{\frac{1}{2}m}} \cdot \frac{s^{2\left(\frac{1}{2}m-1\right)} e^{-\left(\frac{1}{2}ms^2-1\right)/\sigma^2}}{e^{2\left(3+\frac{1}{2}m\right)}} \tag{125}$$

The denominator is

$$f(s^2) = \frac{m^{\frac{1}{2}m} s^{2\left(\frac{1}{2}m-1\right)}}{\mathbf{\Gamma}\left(\frac{1}{2}m\right) 2^{\frac{1}{2}m}} \int_0^\infty \frac{e^{-\frac{1}{2}(ms^2+1)/\sigma^2} d\sigma^2}{\sigma^{2\left(3+\frac{1}{2}m\right)}} = \frac{m^{\frac{1}{2}m} s^{2\left(\frac{1}{2}m-1\right)} \mathbf{\Gamma}\left(2+\frac{1}{2}m\right)}{\mathbf{\Gamma}\left(\frac{1}{2}m\right) 2^{\frac{1}{2}m} \left(\frac{1}{2}ms^2+1\right)^{2+\frac{1}{2}m}}. \tag{126}$$

The resulting posterior distribution is

$$\pi(\sigma^2|s^2) = \frac{(\sigma^2)^{-(a+1)}e^{-1/b\sigma^2}}{\Gamma(a)b^a} \tag{127}$$

with $a = 2 + \frac{1}{2}m$ and $b = \frac{1}{\frac{1}{2}ms^2+1}$.

Notice that we again obtain an inverted gamma distribution. When this situation arises where the prior and posterior distributions come from the same family, we have a *conjugate prior distribution*.

The Bayes rule can be the mean, median, or mode of the posterior distribution.

We now illustrate the result in each of these three cases.

For the mean of (127) and the Bayes rule, we have that

$$\hat{\sigma}_B^2 = E(\sigma^2|s^2) = \frac{(n-1)s^2+2}{n+1}. \tag{128}$$

Its mean and variance are

$$E(\hat{\sigma}_B^2) = \frac{(n-1)\sigma^2+2}{n+1} \quad \text{and} \quad \text{var}(\hat{\sigma}_B^2) = \frac{2(n-1)\sigma^4}{(n+1)^2}. \tag{129}$$

The reader can show that (128) is a biased estimator for $\sigma^2$ with a smaller variance and under certain conditions, smaller mean square error than $s^2$ (Exercise 8).

To estimate the median, it is necessary for a specific sample to solve the equation

$$\int_0^{med} \frac{(\sigma^2)^{-(a+1)}e^{-1/b\sigma^2}}{\mathbf{\Gamma}(a)b^a} d\sigma^2 = 0.5 \tag{130}$$

with $a$ and $b$ from (127) numerically for median.

To find the mode of the posterior distribution, we differentiate the natural logarithm of the posterior distribution, set it equal to zero and solve for $\hat{\sigma}^2$. From (127),

$$\log \pi(\sigma^2|s^2) = -(a+1)\log\sigma^2 - \frac{1}{b\sigma^2}. \tag{131}$$

Differentiating with respect to $\sigma^2$ and setting the result equal to zero, we have that

$$-\frac{a+1}{\sigma^2} + \frac{1}{b(\sigma^2)^2} = 0 \tag{132}$$

and the Bayes estimate is

$$\hat{\sigma}^2 = \frac{1}{(a+1)b^2}. \tag{133}$$

□

***(ii) The Balanced One-Way Random Analysis of Variance Model*** We use the restricted likelihood estimator from (79) with $\lambda = \sigma_e^2 + n\sigma_\alpha^2$,

$$L = \frac{\exp -\frac{1}{2}\left[\dfrac{\text{SSE}}{\sigma_e^2} + \dfrac{\text{SSA}}{(\sigma_e^2 + n\sigma_\alpha^2)}\right]}{(2\pi)^{\frac{1}{2}(an-1)}(\sigma_e^2)^{\frac{1}{2}a(n-1)}(\sigma_e^2 + n\sigma_\alpha^2)^{\frac{1}{2}(a-1)}(an)^{\frac{1}{2}}}. \tag{134}$$

together with the prior distribution of the variance components, a product of inverted gammas

$$\pi(\sigma_e^2, \sigma_\alpha^2) = k\frac{e^{-1/q\sigma_e^2}}{\sigma_e^{2(p+1)}} \cdot \frac{e^{-1/b\sigma_\alpha^2}}{\sigma_\alpha^{2(c+1)}}. \tag{135}$$

The constant $k$ is a function of $c$, $b$, $p$, and $q$ such that $\int_0^\infty \int_0^\infty \pi(\sigma_e^2, \sigma_\alpha^2) d\sigma_e^2 d\sigma_\alpha^2 = 1$. Taking the product of (134) and (135), we shall find its log, find partial derivatives with respect to $\sigma_e^2$ and $\sigma_\alpha^2$ and set them equal to zero. We do not need the denominator of the posterior distribution because it will be a constant function of $\sigma_e^2$ and $\sigma_\alpha^2$. On doing this, we find that we cannot obtain explicit solutions for $\hat{\sigma}_e^2$ and $\hat{\sigma}_\alpha^2$. Let $\hat{\delta} = \hat{\sigma}_e^2/(\hat{\sigma}_e^2 + n\hat{\sigma}_\alpha^2)$ and $\hat{\eta} = \hat{\sigma}_a^2/(\hat{\sigma}_e^2 + n\hat{\sigma}_a^2)$. After some algebraic manipulation, we obtain the equations

$$\hat{\sigma}_e^2 = \frac{\text{SSE}/2 + 1/q + \text{SSA}\hat{\delta}^2/2}{a(n-1)/2 + p + 1 + (a-1)\delta/2} \tag{136a}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{n\text{SSA}\hat{\eta}^2/2 + 1/b}{(a-1)n\hat{\eta}/2 + c + 1}. \tag{136b}$$

For a specific problem, we would have to solve these equations numerically for the estimates of the variance components.

## 11. EXERCISES

**1** Dudewicz and Bishop (1981) describe an experiment that investigates the effects of four bleaching chemicals on pulp brightness. The four chemicals were selected from a large population of potential bleaching agents. The data are as follows.

| Chemical | | Pulp | Brightness | | |
|---|---|---|---|---|---|
| 1 | 77.199 | 74.466 | 92.746 | 76.208 | 82.876 |
| 2 | 80.522 | 79.306 | 81.914 | 80.346 | 73.385 |
| 3 | 79.417 | 78.017 | 91.596 | 80.802 | 80.626 |
| 4 | 78.001 | 78.358 | 77.544 | 77.364 | 77.386 |



(Data taken from Montgomery (2005), problem 13-6, p. 522. Reproduced with kind permission of John Wiley and Sons.)

**(a)** Do the analysis of variance to determine whether there is significant variability amongst the chemicals.

**(b)** Estimate the variance components using

**(1)** The analysis of variance method.

**(2)** Finding the Maximum Likelihood estimator.

**(3)** The Bayes Estimator

Use $q = 1$, $b = 2$, $p = 10$, and $c = 6$.

**(c)** Find the confidence intervals in Table 9.14 when $\alpha = .10$

**2** An experiment was performed to investigate the capability of a measurement system.

Five parts were selected at random and two randomly selected operators measured each part twice.

The tests were made in random order and the following data resulted.

| Operator | 1 | | 2 | |
|---|---|---|---|---|
| | Measurements | | Measurements | |
| Parts | 1 | 2 | 1 | 2 |
| 1 | 50 | 50 | 50 | 51 |
| 2 | 52 | 51 | 51 | 51 |
| 3 | 53 | 50 | 54 | 51 |
| 4 | 49 | 50 | 48 | 51 |
| 5 | 48 | 48 | 48 | 48 |

**(a)** Do the analysis of variance and determine

**(1)** Whether there is significant variability amongst the operators.

**(2)** Whether there is significant variability amongst the parts.

**(3)** Whether the variance component due to interaction is significant.

**(b)** Estimate the variance components

**(1)** Using the ANOVA method.

**(2)** By finding maximum likelihood estimates.

**3** Repeat Exercise 2 assuming there are two specific operators with the parts being selected at random. Test the appropriate hypotheses and estimate the variance components again by the analysis of variance method and the maximum likelihood method.

**4** In a machine shop, 10 people each take one measurement of the length of a 10-inch rod. The measurements obtained from lowest to highest to the nearest thousandth of an inch are 9.339, 9.494, 9.636, 9.682, 9.885, 9.907, 10.101, 10.182, 10.198, and 10.463.

Find the three Bayes estimates of Example 18.

**5** **(a)** Estimate the variance components in Example 9 using the analysis of variance method.

**(b)** Show the equivalence of the quadratic forms

$$Q = \frac{1}{2}\begin{bmatrix}\beta_1 & \beta_2 & \beta_3\end{bmatrix}\begin{bmatrix}20 & -10 & -10\\ -10 & 20 & -10\\ -10 & -10 & 20\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$$

and

$$Q = 15\sum_{j=1}^{3}(\beta_j - \bar{\beta}_\cdot)^2.$$

**6** Consider the following three samples from three populations with mean 10 and different variances.

| 1 | 2 | 3 |
|---|---|---|
| 8.0887 | 11.7611 | 10.2285 |
| 10.4158 | 12.4821 | 10.3067 |
| 10.3684 | 9.4075 | 9.9038 |
| 9.5483 | 13.5566 | 7.0905 |
| 8.9411 | 9.1150 | 9.5776 |
| 8.7716 | 11.0542 | 13.1282 |
| 10.0032 | | 11.1194 |
| 8.8426 | | |

Using equation (112), find the MIVQUE estimates for each sample and compare them to the usual sample variance.

**7** Verify the result in equation (126) and hence establish that the posterior distribution is given by (127). [*Hint:* Let $u = (ms^2 + 1)/(2\sigma^2)$.]

**8** **(a)** Verify the formulae for the mean and variance of (128) that is given in (129).

**(b)** Obtain conditions for the mean square error of (128) to be less than that of $s^2$.

Exercises 9–12 depend on the use of the direct product that we define below.

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be a $p \times q$ matrix. The Kronecker product (also called the direct product) is given by

$$\mathbf{A} \otimes \mathbf{B} = a_{ij}\mathbf{B}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$$

Some useful properties of direct product include:

**(i)** Assume that matrices $\mathbf{A}$ and $\mathbf{B}$ have the same size. Then,

**(a)** $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$

**(b)** $\mathbf{C} \otimes (\mathbf{A} + \mathbf{B}) = \mathbf{C} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{B}$

**(ii)** Assuming $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ have appropriate dimensions so that $\mathbf{AC}$ and $\mathbf{BD}$ are defined,

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}).$$

**(iii)**For two non-singular matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

**(iv)**The transpose $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$.

**(v)** $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}\mathbf{A}\ \text{tr}\mathbf{B}$.

For more details, see Gruber (2014).

**9** **(a)** Show that equation (45) may be written as

$$\mathbf{V} = \mathbf{I}_a \otimes (\sigma_e^2 \mathbf{I}_n + \sigma_\alpha^2 \mathbf{J}_n).$$

**(b)** Show that for the balanced one-way analysis of variance model, equation (46) may be written as

$$\begin{aligned} \text{SSA} &= \mathbf{y}' \left( \frac{1}{n}(\mathbf{I}_a \otimes \mathbf{J}_n) - \frac{1}{na}(\mathbf{J}_a \otimes \mathbf{J}_n) \right) \mathbf{y} \\ &= \mathbf{y}' \left( \left( \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a \right) \otimes \frac{1}{n}\mathbf{J}_n \right) \mathbf{y} \end{aligned}$$

**(c)** Show that

$$\text{SST}_\text{m} = \mathbf{y}' \left( \mathbf{I}_a \otimes \mathbf{I}_n - \frac{1}{an}(\mathbf{J}_a \otimes \mathbf{J}_n) \right) \mathbf{y}$$

and hence

$$\text{SSE} = \mathbf{y}' \left( \mathbf{I}_a \otimes \left( \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n \right) \right) \mathbf{y}.$$

**(d)** Using $E(\mathbf{yy}') = \mathbf{V}$ and $\mathbf{y}'\mathbf{My} = \text{tr}(\mathbf{Myy}')$ show that

**(i)** $E(\text{MSA}) = n\sigma_\alpha^2 + \sigma_e^2$

**(ii)** $E(\text{MSE}) = \sigma_e^2$

**10** Use Theorems 5 and 7 of Chapter 2 to show that $\text{SSA}/(\sigma_e^2 + n\sigma_\alpha^2)$ and $\text{SSE}/\sigma_e^2$ of Table 9.11 in the form of Exercise 7 are distributed independently as central chi-square random variables.

**11** Consider a two-way random model where one factor is nested within the other. Both factors are random. The linear model may be written in the form

$$\mathbf{Y} = (\mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c)\alpha + (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c)\beta + \mathbf{e},$$

where there are $c$ replications.

We can show that

$$\text{SSA} = \mathbf{y}'\left(\left(\mathbf{I}_a - \frac{1}{a}\mathbf{J}_a\right) \otimes \frac{\mathbf{J}_b}{b} \otimes \frac{\mathbf{J}_c}{c}\right)\mathbf{y},$$

$$\text{SSB(A)} = \mathbf{y}'\left(\mathbf{I}_a \otimes \left(\mathbf{I}_b - \frac{1}{b}\mathbf{J}_b\right) \otimes \frac{1}{c}\mathbf{J}_c\right),$$

and

$$\text{SSE} = \mathbf{y}'\left(\mathbf{I}_a \otimes \mathbf{I}_b \otimes \left(\mathbf{I}_c - \frac{1}{c}\mathbf{J}_c\right)\right)\mathbf{y}.$$

**(a)** How many degrees of freedom are associated with each sum of squares.

**(b)** Find the variance of $\mathbf{Y}$ assuming the $\alpha, \beta(\alpha)$ and $\mathbf{e}$ are independent and that their means are zero and that $\alpha \sim (0, \sigma_\alpha^2\mathbf{I}_a), \beta(\alpha) \sim (0, \sigma_{\beta(\alpha)}^2\mathbf{I}_b), \mathbf{e} \sim (0, \sigma_e^2\mathbf{I}_{abc})$.

**(c)** Derive the expected mean squares.

**(d)** Give the analysis of variance estimates of the variance component estimators in terms of MSA, MSB(A), and MSE.

Use the rules for expected mean squares to do Exercises 12, 13, and 15.

**12** For each of the situations below, consider a model for balanced data, more than one replication and all of the possible interactions. Formulate the analysis of variance table as is done in Table 9.3 and give the expected mean squares.

**(a)** Factors $A$, $B$, and $C$ where $C$ is nested within $B$ and all of the factors are random.

**(b)** Factors $A$ and $B$ where factor $C$ is nested within the $\boldsymbol{AB}$ subclasses and $\boldsymbol{D}$ is nested within $\boldsymbol{C}$. Give the expected mean squares when (i) all factors are random; (ii) the model is mixed, where $\boldsymbol{A}$ is a fixed-effects factor, the other factors being random; (iii) the model is mixed, where the fixed factors are $\boldsymbol{A}$ and $\boldsymbol{B}$.

**(c)** Factors $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{D}$ and factor $\boldsymbol{C}$ within $\boldsymbol{AB}$ for the same cases as in part (b).

**13** Consider the following model. It is an example of a split plot design.

$$y_{ijk} = \mu + \alpha_i + \rho_j + \delta_{ij} + \beta_k + \theta_{ik} + e_{ijk}, i = 1, 2, j = 1, 2, 3, 4, k = 1, 2, 3$$

Give the expected mean squares for each of the following cases:

**(a)** Random model,

**(b)** Mixed model $\rho$'s and $\delta's$ random,

**(c)** Mixed model only the $\beta$'s fixed,

**(d)** Mixed model only the $\alpha$'s fixed.

**14** Show that $F = Q/s\hat{\sigma}^2$ as used in earlier chapters (e.g., equation (21) of Chapter 6) is distributed as $F'\{s, N - r, [E(Q) - s\sigma^2]/2\sigma^2\}$.

**15** For each of the following models, obtain the analysis of variance estimators of the variance components in terms of MSA, MSB, etc. Assume the models are random.

**(a)** $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, i = 1, 2, \ldots, a, j = 1, 2, \ldots, b, k = 1, 2, \ldots, n$

**(b)** $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, i = 1, 2, \ldots, a, j = 1, 2, \ldots, b, k = 1, 2, \ldots, n$

**16** Consider the nested model in the form

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c & \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_c & \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_c \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta \end{bmatrix} + e$$

**(a)** Find the normal equations.

**(b)** Solve the normal equations together with the constraints $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_{ij} = 0, 1 \le i \le a$.

**(c)** Show that

**(1)** $\hat{\mu} + \hat{\alpha}_i = \bar{y}_{i..}, 1 \le i \le a$.

**(2)** $\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_{ij} = \bar{y}_{ij.}, 1 \le i \le a, 1 \le j \le b$.

# 10

# METHODS OF ESTIMATING VARIANCE COMPONENTS FROM UNBALANCED DATA

The main focus of Chapter 9 was the estimation of variance components for balanced data by the analysis of variance method. Several other methods that were not discussed there will be presented in this chapter. These methods will be presented largely in general terms. They will be illustrated by means of the one-way and the two-way crossed classifications. Most of the illustrations are of individual aspects of the methods and not of complete analyses. The objective of this chapter is to describe methodology without the clutter of lengthy details of specific cases. This should enable the reader to direct his/her attention to basic procedures instead of being diverted to their numerous details in individual applications. Specific results are available in Chapter 11 posted on the web page: www.wiley.com\go\Searle\LinearModels2E. There, we present these results in full detail with little or no discussion of the methodology. Therefore, the present chapter is a chronicle of the various methods. The web page presents a catalogue of the available consequences of applying these methods to specific cases.

## 1. EXPECTATIONS OF QUADRATIC FORMS

The analysis of variance method of estimating quadratic components from balanced data is based on equating mean squares of analyses of variance to their expected values. This method is well-defined for balanced data because there is only one analysis of variance for any particular model. For example, the only analysis of variance for the balanced two-way classification model with interaction is that of Table 7.9 or equivalently that of Table 9.5 However, for unbalanced data for that

*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

same model, there are two analyses of variance. They are given in the two parts of Table 7.8. One is for fitting $\alpha$ before $\beta$. The other is for fitting $\beta$ before $\alpha$. This is so in general. There can be several, perhaps many ways, or partitioning a sum of squares. On the face of it, there are no criteria for choosing any one of these partitionings over the others when it comes to using one of them for estimating variance components. We shall return to this matter later. For the moment, we only notice that for unbalanced data, there is no uniquely "obvious" sets of sums of squares of quadratic forms in the observations that can be optimally used for estimating variance components. Instead, there are a variety of quadratic forms that can be used, each of them in the method of equating observed quadratic forms to their expected values. Therefore, we begin by considering the expected value of the general quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$.[1]

As usual, we take the general linear model to be

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is $N \times 1$ ($N$ observations). For the sake of generality,

$$\text{var}(\mathbf{y}) = \mathbf{V}.$$

Then, from Theorem 4 of Section 5a of Chapter 2, the expected value of the quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = \text{tr}(\mathbf{Q}\mathbf{V}) + E(\mathbf{y}')\mathbf{Q}E(\mathbf{y}). \tag{2}$$

We view equation (2) in terms of the model (1) for three cases: a fixed-effect model, a mixed model, and a random model.

In each of the three cases, $\mathbf{b}$ represents all of the effects of the model. In addition, in each model, $\text{E}(\mathbf{e}) = \mathbf{0}$, so that $\text{var}(\mathbf{e})$ is $E(\mathbf{e}\mathbf{e}') = \sigma_e^2\mathbf{I}$. Furthermore, when $\mathbf{b}$ is a vector of fixed effects, $E(\mathbf{b}\mathbf{e}') = \mathbf{b}E(\mathbf{e}') = \mathbf{0}$. When $\mathbf{b}$ includes elements that are random effects, we assume they have zero means, and zero covariance with all the elements of $\mathbf{e}$. Thus, at all times, $E(\mathbf{b}\mathbf{e}') = E(\mathbf{e}\mathbf{b}') = \mathbf{0}$.

### a. Fixed-Effects Models

In the usual fixed-effects model, $\mathbf{b}$ is a vector of fixed effects with $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$ and $\mathbf{V} = \sigma_e^2\mathbf{I}_N$. Then, (2) becomes

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} + \sigma_e^2\text{tr}(\mathbf{Q}). \tag{3}$$

Two well-known applications of (3) are $\mathbf{Q} = \mathbf{I}_N$ and $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$. We illustrate equation (3) for these two cases in Examples 1 and 2 below.

[1] The matrix $\mathbf{Q}$ used here is not to be confused with the scalar $Q$ used earlier for the numerator sum of squares in hypothesis testing.

**Example 1 The Case Where $\mathbf{Q} = \mathbf{I}_N$** In this case, we have that

$$E(\mathbf{y}'\mathbf{y}) = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} + N\sigma_e^2. \qquad \square$$

**Example 2 The Case Where $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$** When $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$, we have that $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is the reduction in sum of squares $R(\mathbf{b})$. This gives

$$\begin{aligned} E(R(\mathbf{b})) &= \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} + \sigma_e^2 \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'] \\ &= \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} + \sigma_e^2 r(\mathbf{X}), \end{aligned}$$

because $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ is idempotent and has the same rank as $\mathbf{X}$ (see Theorem 10 of Chapter 1). Hence,

$$E[\mathbf{y}'\mathbf{y} - R(\mathbf{b})] = [N - r(\mathbf{X})]\sigma_e^2.$$

This is the familiar result for a residual sum of squares (see Section 2e of Chapter 5). $\square$

### b. Mixed Models

In a mixed model, we partition $\mathbf{b}'$ as

$$\mathbf{b}' = \left[\mathbf{b}_1' \ \mathbf{b}_A' \ \mathbf{b}_B' \cdots \mathbf{b}_K'\right], \tag{4}$$

where $\mathbf{b}_1$ contains all the fixed effects of the model (including the mean $\mu$). The other $\mathbf{b}$'s each represent a set of random effects for the factors $A$, $B$, $C, \ldots, K$, respectively. Although this notation only uses single subscripts, it does not exclude interaction effects and/or nested-factor effects. We consider them merely as factors, each identified by a single letter rather than the letters of the corresponding main effects. For example, the $AB$-interaction effects might be in the vector $\mathbf{b}_{\boldsymbol{G}}$.

The model (1) is written in terms of (4) as

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_A\mathbf{b}_A + \mathbf{X}_B\mathbf{b}_B + \cdots \mathbf{X}_K\mathbf{b}_K + e,$$

that is, as

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \sum_{\theta=A}^{K} \mathbf{X}_\theta\mathbf{b}_\theta + \mathbf{e}. \tag{5}$$

The matrix $\mathbf{X}$ has been partitioned conformably for the product $\mathbf{Xb}$. In the summation, $\theta$ takes the values $A, B, \ldots, K$. For the random effects, we make two initial assumptions.

(i) They have zero means.
(ii) The effects of each random factor have zero covariance with those of every other factor.

Thus, we write $E(\mathbf{b}_\theta) = \mathbf{0}$ and obtain from (5),

$$E(\mathbf{y}) = \mathbf{X}_1\mathbf{b}_1 \tag{6}$$

and

$$\mathbf{V} = \text{var}(\mathbf{y}) = \sum_{\theta=A}^{K} \mathbf{X}_\theta \text{var}(\mathbf{b}_\theta)\mathbf{X}'_\theta + \sigma_e^2\mathbf{I}_N, \tag{7}$$

where $\mathbf{I}_N$ is an identity matrix of order $N$, and $\text{var}(\mathbf{b}_\theta)$ is the covariance matrix of the random effects of the $\theta$-factor. We usually assume that these effects are uncorrelated, with uniform variance $\sigma_\theta^2$. As a result,

$$\text{var}(\mathbf{b}_\theta) = \sigma_\theta^2\mathbf{I}_{N_\theta} \quad \text{for} \quad \theta = A, B, \ldots, K, \tag{8}$$

there being $N_\theta$ different effects of the $\theta$-factor in the data, meaning $N_\theta$ levels of that factor. Thus in (7), we have that

$$\mathbf{V} = \sum_{\theta=A}^{K} \mathbf{X}_\theta\mathbf{X}'_\theta\sigma_\theta^2 + \sigma_e^2\mathbf{I}_N. \tag{9}$$

Hence, from (6) and (9), the expectation of the quadratic form in (2) is

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = (\mathbf{X}_1\mathbf{b}_1)'\mathbf{Q}\mathbf{X}_1\mathbf{b}_1 + \sum_{\theta=A}^{K} \sigma_\theta^2 \text{tr}(\mathbf{Q}\mathbf{X}_\theta\mathbf{X}'_\theta) + \sigma_e^2\text{tr}(\mathbf{Q}). \tag{10}$$

### c. Random-Effects Models

We take all effects in a random model to be random except for $\mu$, the general mean. Therefore, we can use the expression in (10) just developed for $E(\mathbf{y}'\mathbf{Q}\mathbf{y})$ for the mixed model for the random model, by letting $\mathbf{b}_1$ be the scalar $\mu$ and $\mathbf{X_1}$ be a vector of 1's denoted by $\mathbf{1}$. Thus, we have for the random model,

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = \mu^2\mathbf{1}'\mathbf{Q}\mathbf{1} + \sum_{\theta=A}^{K} \sigma_\theta^2 \text{tr}(\mathbf{Q}\mathbf{X}_\theta\mathbf{X}'_\theta) + \sigma_e^2\text{tr}(\mathbf{Q}). \tag{11}$$

### d. Applications

Applying these general results to particular models involves partitioning $\mathbf{b}$ into subvectors, each of which contains effects pertaining to all levels of one complete classification (or interaction of classifications) involved in the linear model. In this way, expressions (3), (10), and (11) represent the general results for the fixed, mixed,

and random models, respectively. With their aid, expectations of quadratic forms can be readily obtained for any of the three models. For example, suppose we had

$$y_{ijkh} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ik} + e_{ijkh}.$$

We can write this in vector form as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_A\mathbf{b}_A + \mathbf{X}_B\mathbf{b}_B + \mathbf{X}_C\mathbf{b}_C + \mathbf{X}_D\mathbf{b}_D + \mathbf{e},$$

where $\mathbf{b}_A$ is the vector of $\alpha$-effects, $\mathbf{b}_B$ is the vector of the $\beta$'s, and $\mathbf{b}_C$ and $\mathbf{b}_D$ are vectors of the $\gamma$- and $\delta$-terms, respectively. In this way, we can apply the results in (3), (10), and (11) to find expectations of any quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$ of the observations $\mathbf{y}$.

## 2. ANALYSIS OF VARIANCE METHOD (HENDERSON'S METHOD 1)

The analysis of variance method with balanced data consists of equating mean squares to their expected values. We use essentially the same procedure for unbalanced data.

We begin by discussing the method in terms of an example, the two-way classification model. This is not the simplest example that we can use. However, it illustrates facets of the method that could not be demonstrated with a simpler one. We shall give many details of deriving estimators for the two-way classification model but we shall not give the complete results. These will be available on the web page (Chapter 11). In this chapter, we give just those details necessary for illustrating the method and its various aspects.

### a. Model and Notation

The model for the two-way classification with interaction is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \tag{12}$$

We have that $y_{ijk}$ is the $k$th observation in the $i$th level of the $A$-factor and the $j$th level of the $B$-factor where $i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b$, and $k = 1, 2, \ldots, n_{ij}$ with $s$ of the $n_{ij}$-values being non-zero. Section 2a of Chapter 7 gives a complete description of the fixed-effects case of the model. In the random model, which we now consider, we assume that the $\alpha_i$'s, $\beta_j$'s, and $\gamma_{ij}$'s are all random with zero means and variances $\sigma^2_\alpha\mathbf{I}_a$, $\sigma^2_\beta\mathbf{I}_b$, and $\sigma^2_\gamma\mathbf{I}_s$, respectively. This means, for example, that

$$E(\alpha_i) = 0, \quad E\left(\alpha_i^2\right) = \sigma^2_\alpha, \quad \text{and} \quad E(\alpha_i\alpha_{i'}) = 0 \quad \text{for} \quad i \neq i', \tag{13}$$

with similar results for the $\beta$'s and the $\gamma$'s. In addition, we assume that all the covariances between pairs of non-identical random variables are zero. The $e$-terms follow the usual prescription: $E(\mathbf{e}) = \mathbf{0}$, $\text{var}(\mathbf{e}) = \sigma^2_e\mathbf{I}_N$ and the covariance of every $e$ with every random effect is zero.

### b. Analogous Sums of Squares

Table 9.5 shows the analysis of variance for balanced data. It contains a term

$$\text{SSA} = bn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn}, \tag{14}$$

where the bar and the dot notation of totals and means is the same as defined in Section 2 of Chapter 7. For unbalanced data the term analogous to (14) is

$$\text{SSA} = \sum_{i=1}^{a} \frac{y_{i..}^2}{n_{i.}} - \frac{y_{...}^2}{n_{..}}. \tag{15}$$

This is one of the terms used for estimating variance components by the analysis of variance method for unbalanced data. In a similar manner, the other terms are

$$\text{SSB} = \sum_{j=1}^{b} \frac{y_{.j.}^2}{n_{ij}} - \frac{y_{...}^2}{n_{..}}, \tag{16}$$

$$\text{SSAB} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{y_{ij.}^2}{n_{ij}} - \sum_{i=1}^{a} \frac{y_{i..}^2}{n_{i.}} - \sum_{j=1}^{b} \frac{y_{.j.}^2}{n_{.j}} + \frac{y_{...}^2}{n_{..}} \tag{17}$$

and

$$\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}^2 - \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{y_{ij.}^2}{n_{ij}}. \tag{18}$$

The analysis of variance method of variance component estimation for unbalanced data then involves equating (15)–(18) to their expected values. Before considering the derivation of these expected values, we need to make some comments about these SS-terms.

**(*i*) *Empty Cells.*** Since $n_{ij}$ is the number of observations in a cell, it can, as we have seen, be zero. Therefore, the summations in SSAB and SSE that involve $n_{ij}$ in the denominator are therefore defined only for the $(i, j)$ combinations for which $n_{ij}$ is non-zero. That means we sum over only those $s$ cells that have observations in them. This removes the possibility of zero denominators.

**(*ii*) *Balanced Data.*** When the data are balanced, that is, $n_{ij} = \text{n}$ for all $i$ and $j$, then (15) reduces to (14). In a like manner, (16), (17), and (18) reduce to the corresponding analysis of variance sums of squares for balanced data shown in Table 9.5.

**(*iii*) *A Negative "Sum of Squares."*** Equations (15)–(18) have been established solely by analogy with the analysis of variance of balanced data. In general, not all such analogous expressions are sums of squares. For example, SSAB is not always

positive (see Exercise 1 of Chapter 2) and so it is not a sum of squares. We might therefore refer to (15)–(18) and their counterparts in more complicated models as *analogues to sums of squares* in recognition of the fact that the formulae are analogous to sums of squares in the balanced case but not necessarily sums of squares. We could refer to the method as the *analogous analysis of variance method.* However, it is conventionally called the analysis of variance method, or Henderson's method 1, after Henderson (1953).

**(*iv*) *Uncorrected Sums of Squares.*** In light of the fact that, in general, the SS-terms are not sums of squares, we deal with them in terms of uncorrected sums of squares. These are denoted by $T$'s as introduced for balanced data in equation (39) of Section 8 of Chapter 9. Thus for the SS-terms of (15)–(18), we define

$$T_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{n_{i.}} \quad \text{and} \quad T_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{n_{.j}}, \tag{19a}$$

$$T_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{y_{ij.}^2}{n_{ij}} \quad \text{and} \quad T_\mu = \frac{y_{...}}{n_{..}} \tag{19b}$$

with

$$T_0 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}^2. \tag{19c}$$

Apart from $T_\mu$ for the correction factor for the mean and $T_0$ for the total sum of squares of all observations, the subscript of a $T$ denotes the factor it applies to and provides easy recognition of the calculating required. For example,

$$T_A = \sum_{\substack{\text{levels of} \\ A\text{-factor}}} \frac{(\text{total } y \text{ for a level of the } A\text{-factor})^2}{\text{number of observations in that total}}. \tag{20}$$

Similarly $T_{AB}$ is calculated by an expression similar to (20) only with "$A$-factor" replaced by "$AB$-factor." With the $T$'s of (19) the SS-terms in (15)–(18) are

$$\begin{aligned} \text{SSA} &= T_A - T_\mu & \text{and} \quad \text{SSB} &= T_B - T_\mu, \\ \text{SSAB} &= T_{AB} - T_A - T_B + T_\mu & \text{and} \quad \text{SSE} &= T_0 - T_{AB}. \end{aligned} \tag{21}$$

In this form, we may handle the SS-terms with relative ease because the $T$'s are positive definite quadratic forms with manageable matrices.

### c. Expectations

We estimate variance components by equating observed values of terms like (15)–(18) to their expected values. We can calculate the observed values from the $T$'s.

We can derive both the expected values of terms like (15)–(18) (these are quadratic forms) and the expected values of the $T$ from Theorem 4 of Chapter 2. However, the "brute force" method illustrated for balanced data in Section 7 of Chapter 9 is probably no more lengthy than using the theorem. This is especially true when simplifications arising from the model are fully utilized. We therefore illustrate by deriving $E(\text{SSA}) = E(T_A) - E(T_\mu)$ and then give a generalization. The derivation of $E(T_A)$ *in full-length* serves as a guide to deriving expected values of $T$'s generally.

(*i*) ***An Example of a Derivation of the Expectation of a Sum of Squares.*** We obtain

$$E(\text{SSA}) = E(T_A) - E(T_\mu)$$

by substituting the model (12) into $T_A$ and $T_\mu$ of (19) and taking expectations. First, for $T_A$, we have

$$y_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} = n_{i.}\mu^2 + n_{i.}\alpha_i^2 + \sum_{j=1}^{b} n_{ij}\beta_j + \sum_{j=1}^{b} n_{ij}\gamma_{ij} + e_{i..}. \tag{22}$$

Hence, on squaring and expanding the right-hand side of (22) and dividing by $n_{i.}$, we get

$$\frac{y_{i..}^2}{n_{i.}} = \xi + \eta, \tag{23}$$

where

$$\xi = n_{i.}\mu^2 + n_{i.}\alpha_i^2 + \frac{\sum_{j=1}^{b} n_{ij}^2\beta_j^2}{n_{i.}} + \frac{\sum_{j=1}^{b} n_{ij}^2\gamma_{ij}^2}{n_{i.}} + \frac{e_{i..}}{n_{i.}} + \frac{\sum_{j=1}^{b}\sum_{j'\neq j}^{n} n_{ij}n_{ij'}\beta_j\beta_{j'}}{n_{i.}} + \frac{\sum_{j=1}^{b}\sum_{j\neq j'} n_{ij}n_{ij'}\gamma_{ij}\gamma_{ij'}}{n_{i.}}$$

and

$$\eta = 2\left[\mu n_{i.}\alpha_i + \mu\sum_{j=1}^{b} n_{ij}\beta_j + \mu\sum_{j=1}^{b} n_{ij}\gamma_{ij} + \mu e_{i..} + \alpha_i\sum_{j=1}^{b} n_{ij}\beta_j + \alpha_i\sum_{j=1}^{b} n_{ij}\gamma_{ij} + \alpha_i e_{i..}\right]$$
$$+ 2\left[\frac{\left(\sum_{j=1}^{b} n_{ij}\beta_j\right)\left(\sum_{j=1}^{b} n_{ij}\gamma_{ij}\right)}{n_{i.}} + \frac{\left(\sum_{j=1}^{b} n_{ij}\beta_j\right)_i e_{i..}}{n_{i.}} + \frac{\left(\sum_{j=1}^{b} n_{ij}\gamma_{ij}\right)e_{i..}}{n_{i.}}\right].$$

Expression (23) holds true no matter which effects in the model are fixed and which are random.

Consider taking the expected value of (23) under a random model. Products involving $\mu$ go to zero because each other term in such products is a random variable having zero expectation. For example, $E(\mu n_i \alpha_i) = \mu n_i E(\alpha_i) = 0$. Products of random variables also have zero expectation because all covariances and expected values are zero. For example,

$$E\left(\alpha_i \sum_{j=1}^{b} n_{ij}\beta_j\right) = \sum_{j=1}^{b} n_i E(\alpha_i \beta_j) = 0$$

and

$$E(\alpha_i \beta_j) = \text{cov}(\alpha_i \beta_j) + E(\alpha_i)E(\beta_j) = 0.$$

Similarly, we have that

$$\sum_{j=1}^{b} \sum_{j \neq j'}^{b} n_{ij} n_{ij'} E(\beta_j \beta_{j'}) = 0.$$

The only non-zero terms are the expected values of all square terms that, apart from $\mu^2$, become variances. These are the only non-zero terms remaining in $E(y_{i..}^2/n_{i.})$. As a result,

$$E\left(\frac{y_{i..}^2}{n_{i.}}\right) = n_{i.}\mu^2 + n_{i.}\sigma_\alpha^2 + \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\beta^2 + \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\gamma^2 + \sigma_e^2. \tag{24}$$

The last term of (24) is $\sigma_e^2$ because

$$E\left(\frac{e_{i..}^2}{n_{i.}}\right) = \sum_{j=1}^{b} \frac{\sum_{k=1}^{n_{ij}} E\left(e_{ijk}^2\right)}{n_{i.}} = \frac{n_{i.}\sigma_e^2}{n_{i.}} = \sigma_e^2,$$

with the cross products in the $e$'s having zero expectation. Hence, summing (24) gives

$$E(T_A) = E\left(\frac{y_{...}^2}{N}\right) = N\mu^2 + N\sigma_\alpha^2 + \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\beta^2 + \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\gamma^2 + \sigma_e^2. \tag{25}$$

The extended form (23) shows how to derive (24) and (25). It is particularly useful when we come to the case of mixed models where not all the cross product terms have an expected value of zero. For example, see equation (30). However, the consequences of the expected values of the model (e.g., (13)) enable us to go directly from (22) to (25). Thus for $T_\mu$, we write

$$y_{\cdots} = N\mu + \sum_{i=1}^{a} n_{i\cdot}\alpha_i + \sum_{j=1}^{b} n_{\cdot j}\beta_j + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij} + e_{\cdots}.$$

Then, we have that,

$$E(T_\mu) = E\left(\frac{y_{\cdots}^2}{N}\right) = N\mu^2 + \frac{\sum_{i=1}^{a} n_{i\cdot}^2}{N}\sigma_\alpha^2 + \frac{\sum_{j=1}^{b} n_{\cdot j}^2}{N}\sigma_\beta^2 + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}^2}{N}\sigma_\gamma^2 + \sigma_e^2. \quad (26)$$

Hence,

$$\begin{aligned} E(\text{SSA}) &= E(T_A) - E(T_\mu) \\ &= \left(N - \frac{\sum_{i=1}^{a} n_{i\cdot}^2}{N}\right)\sigma_\alpha^2 + \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i\cdot}} - \frac{\sum_{j=1}^{b} n_{\cdot j}^2}{N}\right)\sigma_\beta^2 \\ &\quad + \left(\sum_{i=1}^{a}\frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i\cdot}} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}^2}{N}\right)\sigma_\gamma^2 + (a-1)\sigma_e^2. \end{aligned} \quad (27)$$

In a like manner, we can calculate expected values of SSB and SSAB. Using $E(\text{SSE}) = (N - s)\sigma_e^2$, when we equate the four expected values to their corresponding observed values, we obtain four equations in the four variance components we wish to estimate.

Notice that (27) has a non-zero coefficient for every variance component in the model. For balanced data, the comparable expected value has no term in $\sigma_\beta^2$ (see $E$(MSA) in Table 9.8). However, the coefficient for the term in $\sigma_\beta^2$ does reduce to zero for balanced data. Indeed, when

$$n_{ij} = n, \quad n_{i\cdot} = bn, \quad n_{\cdot j} = an, \quad \text{and} \quad n_{\cdot\cdot} = N = abn, \quad (28)$$

the coefficient of $\sigma_\beta^2$ in (27) is

$$\sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}} - \frac{\sum_{j=1}^{b} n_{.j}^2}{N} = a\left(\frac{bn^2}{bn}\right) - \frac{ba^2n^2}{abn} = an - an = 0.$$

Similarly, the coefficient of $\sigma_\alpha^2$ in (27) becomes

$$N - \frac{\sum_{i=1}^{a} n_{i.}^2}{N} = abn - \frac{ab^2n^2}{abn} = bn(a-1).$$

The coefficient of $\sigma_\gamma^2$ reduces to $n(a-1)$. Hence, for balanced data (27) becomes

$$E(\text{SSA}) = (a-1)\left(bn\sigma_\alpha^2 + n\sigma_\gamma^2 + \sigma_e^2\right)$$

as is implicit in Table 9.8.

**(*ii*) *Mixed Models.*** Suppose that in the two-way classification, the $A$-factor is a fixed-effects factor. Then, the $\alpha_i$'s of the model are fixed effects. Furthermore, the expected values of the SS-terms of (21) differ from their values under the random model. For example, in taking the expected value of (23) to obtain $E(T_A)$, we have, with the $\alpha$'s as fixed effects,

$$\begin{aligned} &E(n_{i.}\alpha_i^2) = n_{i.}\alpha_i^2, \text{ and not } n_{i.}\sigma_\alpha^2 \text{ as in (24);} \\ &E(2\mu n_{i.}\alpha_i) = 2\mu n_{i.}\alpha \text{ and not 0 as in (24).} \end{aligned} \tag{29}$$

Other terms in (23) involving $\alpha_i$ will have zero expectation, just as they did in (24) but now for a different reason. For example, $E(\alpha_i\beta_j) = 0$ in (24) because the $\alpha$'s and $\beta$'s were random variables with zero means and covariances. In the mixed model, $E(\alpha_i\beta_j)$ is still equal to zero. However, this is because $E(\alpha_i\beta_j) = \alpha_i E(\beta_j) = \alpha_i(0) = 0$.

Equations (29) mean that in the mixed model, instead of the terms $N\mu^2 + N\sigma_\alpha^2$,

$$E(T_A) \text{ contains } N\mu^2 + \sum_{i=1}^{a} n_{i.}\alpha_i^2 + 2\mu \sum_{i=1}^{a} n_i\alpha_i. \tag{30}$$

Similarly, we can show that

$$E(T_\mu) \text{ contains } N\mu^2 + \frac{\left(\sum_{i=1}^{a} n_{i.}\alpha_i\right)^2}{N} + 2\mu \sum_{i=1}^{a} n_{i.}\alpha_i. \tag{31}$$

Therefore,

$$E(\text{SSA}) = E(T_A) - E(T_\mu) \text{ contains } \sum_{i=1}^{a} n_{i.}\alpha_i^2 - \frac{\left(\sum_{i=1}^{a} n_{i.}\alpha_i\right)^2}{N} = \theta_1, \text{ say.} \tag{32}$$

Carrying through the same process for SSB shows that

$$E(\text{SSB}) = E(T_B) - E(T_\mu) \text{ contains } \sum_{j=1}^{b} \frac{\left(\sum_{i=1}^{a} n_{ij}\alpha_i\right)^2}{n_{.j}} - \frac{\left(\sum_{i=1}^{a} n_{i.}\alpha_i\right)^2}{N} = \theta_2, \text{ say.} \tag{33}$$

The important thing to notice is that $\theta_1 \neq \theta_2$. Thus, $E(\text{SSA} - \text{SSB})$ is not free of fixed effects in the way that $E(T_A) - E(T_\mu)$ is of $N\mu^2$. This is generally true for mixed models. Expected values of the SS-terms contain functions of the fixed effects that cannot be eliminated by considering linear functions of the terms. Thus, the analysis of variance method cannot be used for mixed models.

We present two possible ways to overcome the above difficulty. However, both are deviants from the true mixed model and must therefore be considered unsatisfactory. The two ways are:

(i) Ignore the fixed effects altogether and eliminate them from the model. What remains then is a model that is completely random for which estimation of the variance components can be made.

(ii) Assume the fixed effects are in fact random, and then treat the model as if it were completely random. In the resulting estimation process, components for the fixed effects will be estimated and can be ignored.

In using either of these possibilities, we deal with random models, for which the estimation process is suitable. However, the variance component estimators will, in both cases, be biased because their expectations under the true, mixed model will not equal the variance components of that model. They will include quadratic functions of the fixed effects. Despite this, if the models that these approximations invoke are in any way acceptable alternatives to the mixed model, then the approximations may be of some use. Furthermore, they utilize the relatively easy arithmetic of the analysis of variance method. This is sometimes advantageous in the face of the greater complexity of other analyses of mixed models (see Section 3).

**(*iii*) *General Results.*** We now develop general rules for obtaining expectations of the *T*-terms in random models. To do so, we write the model as

$$\mathbf{y} = \mu\mathbf{1} + \sum_{\theta=A}^{K} \mathbf{X}_\theta \mathbf{b}_\theta + \mathbf{e}. \tag{34}$$

The model in (34) is the same as that in (5), with $\mathbf{X_1} = \mathbf{1}$ and $\mathbf{b_1}$ taken as the scalar $\mu$. To derive $E(T_A)$ from (20), we define

$$\left.\begin{array}{l} y_{\cdot}(A_i) = \text{total of} \\ n(A_i) = \text{number of} \end{array}\right\} \text{observations in the } i\text{th level of the } A\text{-factor}$$

and have from (20),

$$T_A = \sum_{i=1}^{N_A} \frac{[y_{\cdot}(A_i)]^2}{n(A_i)}. \tag{35}$$

Now, just as in Section 7 of Chapter 6, define $n(A_i, \theta_j)$ as the number of observations in the $i$th level of the $A$-factor and the $j$th level of the $\theta$-factor. In addition, define $b_{\theta_j}$ as the $j$th element of $\mathbf{b}_\theta$, and $e_{\cdot}(A_i)$ as the total of the error terms corresponding to $y_{\cdot}(A_i)$. Then, using (34) in (35) gives

$$T_A = \sum_{i=1}^{N_A} \frac{\left[ n(A_i)\mu + \sum_{\theta=A}^{K} \sum_{j=1}^{N_\theta} n(A_i, \theta_j) b_{\theta_j} + e_{\cdot}(A_i) \right]^2}{n(A_i)} \tag{36}$$

Taking expected values of (36) gives:

(i) a term in $\mu^2$:

$$\sum_{i=1}^{N_A} \frac{[n(A_i)]^2 \mu^2}{n(A_i)} = \mu^2 \sum_{i=1}^{N_A} n(A_i) = N\mu^2; \tag{37}$$

(ii) a term in $\sigma_\theta^2$, for $\theta = A, B, \ldots, K$:

$$k\left(\sigma_\theta^2, T_A\right)\sigma_\theta^2 = \sum_{i=1}^{N_A} \frac{\sum_{i=1}^{N_\theta} [n(A_i, \theta_j)]^2}{n(A_i)} \sigma_\theta^2, \tag{38}$$

where we define $k(\sigma_\theta^2, T_A)$ as the coefficient of $\sigma_\theta^2$ in $E(T_A)$;

(iii) a term in $\sigma_e^2$:

$$\sum_{i=1}^{N_A} \frac{n(A_i)\sigma_e^2}{n(A_i)} = \sigma_e^2 \sum_{i=1}^{N_A} 1 = N_A \sigma_e^2. \tag{39}$$

**Example 3 Derivation of $E(T_a)$ of (25) from (39)** The terms $N\mu^2$ and $N_A\sigma_e^2$ need no demonstration. The others are

$$k\left(\sigma_\alpha^2, T_A\right)\sigma_\alpha^2 = \sum_{i=1}^{a} \frac{[n(\alpha_i, \alpha_i)]^2}{n(\alpha_i)}\sigma_\alpha^2 = \sum_{i=1}^{a} n_{i.}\sigma_\alpha^2 = N\sigma_\alpha^2,$$

$$k\left(\sigma_\beta^2, T_A\right)\sigma_\beta^2 = \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} [n(\alpha_i, \beta_j)]^2}{n(\alpha_i)}\sigma_\beta^2 = \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\beta^2,$$

and

$$k\left(\sigma_\gamma^2, T_A\right)\sigma_\gamma^2 = \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} [n(\alpha_i, \gamma_{ij})]^2}{n(\alpha_i)}\sigma_\gamma^2 = \sum_{i=1}^{a} \frac{\sum_{j=1}^{b} n_{ij}^2}{n_{i.}}\sigma_\gamma^2.$$

Similarly, the terms in $E(T_\mu)$ are, for example, of the form:

$$k\left(\sigma_\alpha^2, T_\mu\right)\sigma_\alpha^2 = \frac{\sum_{i=1}^{a} [n(\alpha_i)]^2}{N}\sigma_\alpha^2 = \frac{\sum_{i=1}^{a} n_{i.}^2}{N}\sigma_\alpha^2$$

as in (26).

**(*iv*) *Calculation by "Synthesis."*** Hartley (1967) developed a method for calculating coefficients of $\sigma^2$'s in terms like $E$(SSA and $E(T_A)$ without first requiring the algebraic form of these coefficients. The method applies to calculating coefficients of the $\sigma^2$'s in expected values of any quadratic form that is homogeneous in the observations in **y**. It requires no distributional properties of the model. He has called it the method of "synthesis." We describe it in terms of calculating $T_A$ of Sub-Section 2c(iii).

Write $T_A$ of (35) as

$$T_A = \sum_{i=1}^{N_A} \frac{[y_{\cdot}(A_i)]^2}{n(A_i)} = \mathbf{y}'\mathbf{Q}_A\mathbf{y} = T_A(\mathbf{y}). \tag{40}$$

Define

$$x(\theta, j) = j\text{th column of } \mathbf{X}_\theta. \tag{41}$$

Then, the method of synthesis derives $k(\sigma_\theta^2, T_A)$ as the coefficient of $\sigma_\theta^2$ in $E(T_A)$, as

$$k\left(\sigma_\theta^2, T_A\right) = \sum_{j=1}^{N_\theta} T_A[\mathbf{x}(\theta, j)]; \tag{42}$$

that is, using each column of $\mathbf{X}_\theta$ as a column of data (all 0's and 1's), calculate $T_A$, and sum the results over all columns of $\mathbf{X}_\theta$. The sum is the coefficient of $\sigma^2_\theta$ in $E(T_A)$, namely $k(\sigma^2_\theta, T_A)$ of (38).

The procedure can be used numerically without recourse to explicit algebraic forms of the coefficients $k(\sigma^2_\theta, T_A)$. Since it applies to any quadratic form in the place of $T_A$, it can also be used directly on the SS-terms. Thus, paraphrasing Hartley's words: we can apply the analysis of variance method to each of the $N_\theta$ columns of $\mathbf{X}_\theta$ used as data. Single out a particular quadratic $f(\mathbf{y})$ over the $N_\theta$ analyses of variance, to obtain $k[\sigma^2_\theta, f(\mathbf{y})]$, the coefficient of $\sigma^2_\theta$ in $E[f(\mathbf{y})]$. Therefore, carrying out $\sum_{\theta=A}^{k} N_\theta$ analyses of variance and summing them appropriately gives all the coefficients of the $\sigma^2$'s in the expected quadratics. Since many of the "observations" in these analyses will be zero, any computer procedure designed for this task should take account of this many-zeroed feature of the "data."

We can show the equivalence of (42) to (38). The $j$th column of $\mathbf{X}_\theta$, namely $\mathbf{x}(\theta, j)$, has $n(\theta_j)$ ones in it and $N - n(\theta_j)$ zeroes. Therefore, using $\mathbf{x}(\theta, j)$ as the vector $\mathbf{y}$ in $y_{\cdot}(A_i)$ of (40) we require the total of the "observations" in $\mathbf{x}(\theta, j)$ that are in the $i$th level of $A$. These observations will consist of $n(A_i, \theta_j)$ ones and $n(A_i) - (n(A_i, \theta_j)$ zeros. Their total is thus $n(A_i, \theta_j)$. Therefore, from (40),

$$T_A[\mathbf{x}(\theta, j)] = \sum_{i=1}^{N_a} \frac{[n(A_i, \theta_j)]^2}{n(A_i)}.$$

Summing this over $j$ as in (42) yields (38).

The method of "synthesis" can be applied to calculating variances of variance component estimators (see Section 2d(iii) following). J. N. K. Rao (1968) extended it to general incidence matrices and to mixed models.

### d. Sampling Variances of Estimators

The analogous sums of squares (in the manner of (14)–(17)) that we use in the analysis of variance method for unbalanced data are the SS-terms. Under normality assumptions, they do not have $\chi^2$-distributions. In addition, they are not distributed independently of each other. The only sum of squares with a known distribution is SSE. It follows a $\chi^2$-distribution in the usual manner. In addition, it has zero covariance with the other SS-terms. Therefore, $\hat{\sigma}^2 = \text{SSE}/(N - s)$ has a similar distribution. The other distributions that are linear functions of the SS-terms have distributions that are unknown. Despite this, under normality assumptions, we can derive the variances of these estimators. Suppose we define

$$\mathbf{c} = \text{vector of SS-terms but not SSE,}$$
$$\sigma^2 = \text{vector of } \sigma^2\text{'s, but not } \sigma^2_e,$$

and

$$\mathbf{f} = \text{vector of "degrees of freedom," the coefficients of } \sigma^2_e \text{ in } E(\mathbf{c}).$$

The vector of SS-terms is therefore $\begin{bmatrix}\mathbf{c}' & \text{SSE}\end{bmatrix}$. Equating this to its expected values yields the variance components estimators. Suppose $\mathbf{P}$ is the matrix of coefficients of variance components (other than $\sigma_e^2$) in $E(\mathbf{c})$. Then, we can write

$$E\begin{bmatrix}\mathbf{c}\\ \text{SSE}\end{bmatrix} = \begin{bmatrix}\mathbf{P} & \mathbf{f}\\ \mathbf{0} & N-s\end{bmatrix}\begin{bmatrix}\sigma^2\\ \sigma_e^2\end{bmatrix}. \tag{43}$$

Equating $\begin{bmatrix}\mathbf{c}\\ \text{SSE}\end{bmatrix}$ to its expected values gives the estimators

$$\hat{\sigma}_e^2 = \frac{\text{SSE}}{(N-s)} \tag{44a}$$

and

$$\hat{\sigma}^2 = \mathbf{P}^{-1}\left(\mathbf{c} - \hat{\sigma}_e^2\mathbf{f}\right). \tag{44b}$$

Expressions (44a) and (44b) provide a means for deriving the variances of the estimators.

**(*i*) *Derivation.*** The distribution of $\text{SSE}/\sigma_e^2$ is $\chi^2_{N-s}$ with variance $2(N-s)$. Thus, from (44a), we have,

$$v\left(\hat{\sigma}_e^2\right) = \frac{2\sigma_e^4}{(N-s)}. \tag{45}$$

Now SSE (and hence $\hat{\sigma}_e^2$) has zero covariance with every element of $\mathbf{c}$, that is, with every other SS-term. Therefore, from (44b),

$$\text{cov}\left(\hat{\sigma}^2, \hat{\sigma}_e^2\right) = -\mathbf{P}^{-1}\mathbf{f}v\left(\hat{\sigma}_e^2\right) \tag{46a}$$

and

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\left[\text{var}(\mathbf{c}) + v\left(\hat{\sigma}_e^2\right)\mathbf{f}\mathbf{f}'\right]\mathbf{P}^{-1'}. \tag{46b}$$

In addition, since the SS-terms are linear functions of the $T$'s, we can with

$$\mathbf{T} = \text{vector of } T\text{'s write } \mathbf{c} = \mathbf{H}\mathbf{t} \tag{47}$$

for some matrix $\mathbf{H}$ (that is quite unrelated to $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$ of previous chapters). For example, in the case of the two-way classification, $\mathbf{H}$ is the matrix of the transformation of the $T$'s to SSA, SSB, and SSAB shown in (21). Hence,

$$\text{var}(\mathbf{c}) = \mathbf{H}\text{var}(t)\mathbf{H}' \tag{48a}$$

and

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\left[\mathbf{H}\text{var}(\mathbf{t})\mathbf{H}' + v\left(\hat{\sigma}_e^2\right)\mathbf{f}\mathbf{f}'\right]\mathbf{P}. \tag{48b}$$

This result is due by Searle (1958). Blischke (1968) utilized this result in the general case. Its application in any particular situation requires obtaining only var(**t**), the variance–covariance matrix of the $T$'s. The matrix **P** is that of the coefficients of the $\sigma^2$'s in the expected values of the SS-terms. The matrix **H** expresses the relationship between the SS-terms and the $T$'s. The vector of the "degrees of freedom" in the SS-terms is **f**. The elements of **f** are the coefficients of $\sigma_e^2$ in the expected values of the SS-terms.

Deriving elements of var(**t**) requires cumbersome algebra. However, the basis of two different methods of doing so is quite straightforward. For both methods, we assume normality, that is, that

$$\mathbf{y} \sim N(\mu\mathbf{1}, \mathbf{V}). \tag{49}$$

For the first method, we show the manner in which $\mu^2$ occurs in the variances and the covariances of the $T$'s. From (40), we can show that

$$\mathbf{Q}_A = \sum_{i=1}^{N_A} + \frac{1}{n(A_i)}\mathbf{J}_{n(A_i)}. \tag{50}$$

This means that $\mathbf{Q}_A$ is a block diagonal matrix of square matrices of order $n(A_i)$ with every element being $1/n(A_i)$. This kind of result applies not just to the $A$-factor but also to every factor $\theta$ of the model (34). For two factors $A$ and $B$, we then have from Chapter 2,

$$v(T_A) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A)^2 + 4\mu^2\mathbf{1}'\mathbf{Q}_A\mathbf{V}\mathbf{Q}_A\mathbf{1}.$$

There is a similar expression for $\text{v}(T_B)$. Furthermore,

$$\text{cov}(T_A, T_B) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A\mathbf{V}\mathbf{Q}_B) + 4\mu^2\mathbf{1}'\mathbf{Q}_A\mathbf{V}\mathbf{Q}_B\mathbf{1}.$$

However, from (50),

$$\mathbf{1}'\mathbf{Q}_A = \mathbf{1}'.$$

The same is true of $\mathbf{Q}_B$. Thus, we have that

$$v(T_A) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A)^2 + 4\mu^2\mathbf{1}'\mathbf{V}\mathbf{1}$$

and that

$$\text{cov}(T_A, T_B) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A\mathbf{V}\mathbf{Q}_B) + 4\mu^2\mathbf{1}'\mathbf{V}\mathbf{1}.$$

Observe that $4\mu^2\mathbf{1}'\mathbf{V}\mathbf{1}$ is part of all the variances and covariances of the $T$'s. However, because in $\mathbf{c} = \mathbf{H}\mathbf{t}$ the $T$'s are used only in terms of the differences between them, the $4\mu^2\mathbf{1}'\mathbf{V}\mathbf{1}$ term in the above expressions can be ignored. This is equivalent to assuming $\mu = 0$. It gives

$$v(T_A) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A)^2 \tag{51}$$

and

$$\text{cov}(T_A, T_B) = 2\text{tr}(\mathbf{V}\mathbf{Q}_A\mathbf{V}\mathbf{Q}_B). \tag{52}$$

From these elements, we can obtain the elements of var($\mathbf{t}$). This has been done for several specific cases. The details are available on the web page (Chapter 11).

The second method is due to Blischke (1966, 1968). He obtains the same elements of var($\mathbf{t}$) by using the fact that for normal variables $u$ and $v$,

$$\text{cov}(u^2, v^2) = 2[\text{cov}(u, v)]^2$$

(See Exercise 17 of Chapter 2). Therefore, since $T_A$ and $T_B$ are weighted sums of squares of normally distributed random variables, their covariance

$$\text{cov}(T_A, T_B) = \text{cov}\left\{\sum_{i=1}^{N_A}\frac{[y_{\cdot}(A_i)]^2}{n(A_i)}, \sum_{j=1}^{N_B}\frac{[y_{\cdot}(B_j)]^2}{n(B_j)}\right\},$$

is, assuming $\mu = 0$,

$$\text{cov}(T_A, T_B) = \sum_{i=1}^{N_A}\sum_{j=1}^{N_B}\frac{2\{\text{cov}[y_{\cdot}(A_i), y_{\cdot}(B_j)]\}^2}{n(A_i)n(B_j)}.$$

A special case of this is

$$\text{var}(T_A) = \sum_{i=1}^{N_A}\frac{2\{\text{var}[y_{\cdot}(A_i)]\}^2}{[n(A_i)]^2} + \sum_{i\neq i'}^{N_A}\frac{2\{\text{cov}[y_{\cdot}(A_i), y_{\cdot}(A_{i'})]\}^2}{n(A_i)n(A_{i'})}.$$

Whether one uses these expressions or their equivalent matrix forms (51) and (52), the ensuing algebra for specific cases is cumbersome and tedious, as is evident from the results listed on the web page (Chapter 11).

One of the difficulties in deriving the matrix var($\mathbf{t}$) is the large number of elements. An $r$-way classification random model, with all interactions (see Exercise 2) involves $2^{r-1}(2^r + 1)$ different elements in var($\mathbf{t}$). Each element is a linear function of the same number of squares and products of variance components. Thus, we need to deal with a square matrix of order $2^{r-1}(2^r + 1)$. For $r = 2$, 3, 4, and 5, this matrix has order 10,

36, 136, and 528, respectively. Its elements for $r = 2$ and $r = 3$ are available on the web page (Chapter 11).

**(*ii*) *Estimation.*** We may use the elements of var(**t**) that we derive from (51) and (52) in (48) to obtain var($\hat{\sigma}^2$). However, the elements of var(**t**) are quadratic functions of the unknown variance components. The problem of estimating var($\hat{\sigma}^2$) therefore remains. A common procedure is to replace the variance components in var($\hat{\sigma}^2$) by their estimates and use the resulting value of var($\hat{\sigma}^2$) as the estimator of var($\hat{\sigma}^2$). As an estimator, this has no known desirable properties other than being relatively easy to compute. Searle (1961) discusses a small numerical example. We can derive unbiased estimators of the variances and covariances of the variance component estimator, that is, of (45), (46), and (48) as follows. First, array (45), (46), and the elements of the upper triangular half of (48) in a vector **v**:

$$\mathbf{v} = \text{vector of variances and covariances of all } \hat{\sigma}^2\text{'s.}$$

For example, in the one-way classification with components $\sigma_\alpha^2$ and $\sigma_e^2$,

$$\mathbf{v}' = \left[ v\left(\hat{\sigma}_\alpha^2\right) \quad v\left(\hat{\sigma}_e^2\right) \quad \operatorname{cov}\left(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2\right) \right]$$

and

$$\gamma' = \left[ \sigma_\alpha^4 \quad \sigma_e^4 \quad \sigma_\alpha^2 \sigma_e^2 \right].$$

Then, because of (45), (46), and (48), every element of **v** is a linear combination of the elements in $\gamma$, and so, for some matrix **A** say

$$\mathbf{v} = \mathrm{A}\gamma \tag{53}$$

With an $r$-way classification random model that has all possible interactions **A** of (53) has an order $2^{r-1}(2^r + 1)$. However, **A** is not the matrix referred to at the end of Section (i), where the different elements of var(**t**) were envisaged as a vector $\mathbf{B}\gamma$, say. In (53), it is $v(\hat{\sigma}_e^2)$ and the elements of $\operatorname{cov}(\hat{\sigma}^2, \hat{\sigma}_e^2)$ and var($\hat{\sigma}^2$) being written as $\mathbf{A}\gamma$. The matrices **A** and **B** have the same order but are not equal.

We derive the unbiased estimator of **v** from (53). First, observe that every variance component estimator of $\hat{\sigma}^2$ of (44) is unbiased. Thus, for example, on writing $\hat{\sigma}_A^4$ for $(\hat{\sigma}_A^2)^2$, we have

$$E\left(\hat{\sigma}_A^4\right) = v\left(\hat{\sigma}_A^2\right) + \sigma_A^4. \tag{54}$$

Similarly,

$$E\left(\hat{\sigma}_A^2 \hat{\sigma}_B^2\right) = \operatorname{cov}\left(\hat{\sigma}_A^2, \hat{\sigma}_B^2\right) + \sigma_A^2 \sigma_B^2. \tag{55}$$

Writing $\hat{\gamma}$ as the vector of squares and products of the $\hat{\sigma}^2$'s corresponding to $\gamma$, we have from (54) and (55) that

$$E(\hat{\gamma}) = \mathbf{v} + \gamma. \tag{56}$$

We will then find that replacing $\gamma$ in (53) by $\hat{\gamma} - \hat{\mathbf{v}}$ and calling the resulting expression $\hat{\mathbf{v}}$ yields $\hat{\mathbf{v}}$ as an unbiased estimator of $\mathbf{v}$; that is,

$$\hat{\mathbf{v}} = \mathbf{A}(\hat{\gamma} - \hat{\mathbf{v}}) \tag{57}$$

gives

$$\hat{\mathbf{v}} = (\mathbf{I} + \mathbf{A})^{-1}\mathbf{A}\hat{\gamma} \tag{58}$$

as an unbiased estimator of $\mathbf{v}$. Utilizing (53) and (56) in taking the expected value of (58) shows that $E(\hat{\mathbf{v}}) = \mathbf{v}$ (see Exercise 6). The elements in $\hat{\mathbf{v}}$ in (58) are therefore unbiased estimators of the variances and covariances of the analysis of variance estimators of the variance components.

Mahamunulu (1963) describes the derivation of $\hat{\mathbf{v}}$ in terms of (57). It consists of replacing every $\sigma_A^4$ term in (53) by $\hat{\sigma}_A^4 - v(\hat{\sigma}_A^2)$ and every $\sigma_A^2\sigma_B^2$ term in (53) by $\hat{\sigma}_A^2\hat{\sigma}_B^2 - \hat{\text{cov}}(\hat{\sigma}_A^2, \hat{\sigma}_B^2)$, and calling the resulting expression $\hat{\mathbf{v}}$. Ahrens (1965) derived the form of the result given in (58).

The nature of (45) ensures that (58) yields

$$\hat{v}\left(\hat{\sigma}_e^2\right) = \frac{2\hat{\sigma}^4}{(N - s + 2)}. \tag{59}$$

We can, of course, also derive (59) directly from (45) by using the counterpart of (54) for $\sigma_e^2$. In the same way, (58) also yields

$$\hat{\text{cov}}\left(\hat{\sigma}^2, \hat{\sigma}_e^2\right) = -\mathbf{Pf}\hat{v}\left(\hat{\sigma}_e^2\right) \tag{60}$$

as an unbiased estimator of (46). The remaining terms in $\hat{\mathbf{v}}$ are unbiased estimators of the elements of $\text{var}(\hat{\sigma}^2)$ of (48).

**Example 4 Estimating Variances of Variance Components in the Unbalanced One-Way Classification Model** The analysis of variance for the one-way classification model is derived in Section 2d of Chapter 6. Denoting $\text{SSR}_\text{m}$ given there as SSA, we have

$$\text{SSA} = \sum_{i=1}^{a} n_i \bar{y}_{i.}^2 - N y_{..}^2, \quad \text{with} \quad E(\text{SSA}) = \left( N - \frac{\sum_{i=1}^{a} n_i^2}{N} \right) \sigma_\alpha^2 + (a - 1)\sigma_e^2$$

and

$$\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a} n_i \bar{y}_{i.}^2, \quad \text{with} \quad E(\text{SSE}) = (N-a)\sigma_e^2.$$

Therefore, from (43), $\mathbf{P}$ and $\mathbf{f}$ are scalars,

$$\mathbf{P} = N - \frac{\sum_{i=1}^{a} n_i^2}{N} \quad \text{and} \quad \mathbf{f} = a - 1. \tag{61}$$

The estimators are

$$\hat{\sigma}_e^2 = \frac{\text{SSE}}{N-a} \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{\text{SSA} - (a-1)\hat{\sigma}_e^2}{N - \sum_{i=1}^{a} n_i^2/N}. \tag{62}$$

The variances and covariances of these estimators are (see Crump (1947) and Searle (1956))

$$\begin{aligned} v\left(\hat{\sigma}_e^2\right) &= k_1\sigma_e^4 \quad \text{for} \quad k_1 = \frac{2}{(N-a)}, \\ \text{cov}\left(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2\right) &= k_2\sigma_e^4 \quad \text{for} \quad k_2 = \frac{-2(a-1)}{[(N-a)(N-S_2/N)]}, \end{aligned} \tag{63}$$

and

$$v\left(\sigma_\alpha^2\right) = k_3\sigma_e^4 + k_4\sigma_e^2\sigma_\alpha^2 + k_5\sigma_\alpha^4$$

with

$$k_3 = \frac{2N^2(N-1)(a-1)}{(N^2-S_2)^2},\ k_4 = \frac{4N}{N^2-S_2},\ \text{and}\ k_5 = \frac{2\left(N^2S_2 + S_2^2 - 2NS_3\right)}{(N^2-S_2)^2},$$

where $S_2 = \sum_{i=1}^{a} n_i^2$ and $S_3 = \sum_{i=1}^{a} n_i^3$. Therefore, (53) is

$$\begin{bmatrix} v\left(\hat{\sigma}_e^2\right) \\ \text{cov}\left(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2\right) \\ v\left(\hat{\sigma}_\alpha^2\right) \end{bmatrix} = \begin{bmatrix} k_1 & 0 & 0 \\ k_2 & 0 & 0 \\ k_3 & k_4 & k_5 \end{bmatrix} \begin{bmatrix} \sigma_e^4 \\ \sigma_\alpha^2\sigma_e^2 \\ \sigma_\alpha^4 \end{bmatrix}.$$

As a result, (58) is

$$\begin{bmatrix} \hat{v}(\hat{\sigma}_e^2) \\ \hat{\text{cov}}(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2) \\ \hat{v}(\hat{\sigma}_\alpha^2) \end{bmatrix} = \begin{bmatrix} 1+k_1 & 0 & 0 \\ k_2 & 1 & 0 \\ k_3 & k_4 & 1+k_5 \end{bmatrix}^{-1} \begin{bmatrix} k_1 & 0 & 0 \\ k_2 & 0 & 0 \\ k_3 & k_4 & k_5 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_e^4 \\ \hat{\sigma}_\alpha^2 \hat{\sigma}_e^2 \\ \hat{\sigma}_\alpha^4 \end{bmatrix}$$
$$= \frac{1}{(1+k_1)(1+k_5)} \begin{bmatrix} k_1(1+k_5) & 0 & 0 \\ k_2(1+k_5) & 0 & 0 \\ k_3 - k_2 k_4 & k_4(1+k_1) & k_5(1+k_1) \end{bmatrix} \begin{bmatrix} \hat{\sigma}_e^4 \\ \hat{\sigma}_\alpha^2 \hat{\sigma}_e^2 \\ \hat{\sigma}_\alpha^4 \end{bmatrix}. \tag{64}$$

From (64),

$$\hat{v}(\hat{\sigma}_e^2) = \frac{k_1(1+k_5)\hat{\sigma}_e^4}{(1+k_1)(1+k_5)} = \frac{k_1 \hat{\sigma}_e^4}{1+k_1} = \frac{2\hat{\sigma}_e^4}{N-a+2}$$

on substituting for $k_1$, a result that is in keeping with (59). Similarly, from (64),

$$\hat{\text{cov}}(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2) = \frac{k_2}{1+k_1}\hat{\sigma}_e^4 = \frac{k_2}{k_1}\left(\frac{k_1 \hat{\sigma}_e^4}{1+k_1}\right) = \frac{k_2}{k_1}\hat{v}(\hat{\sigma}_e^2).$$

This result agrees with (60) because, from (63) and (61), $k_2/k_1 = -\mathbf{P}^{-1}\mathbf{f}$ of (60). □

**Example 5 Numerical Estimates of Variance Components and Their Variances and Covariances for a One-Way Classification Model** Four brands of light bulbs are chosen from a large population of light-bulb brands. The life lengths of samples of light bulbs are given below.

| A | B | C | D |
|---|---|---|---|
| 915 | 1011 | 989 | 1055 |
| 912 | 1001 | 979 | 1048 |
| 903 | 1003 | 1004 | 1061 |
| 893 | 992 | 992 | 1068 |
| 910 | 981 | 1008 | 1053 |
| 890 | 1001 | 1009 | 1063 |
| 879 | 989 | 996 | |
| | 1003 | 998 | |
| | | 997 | |

(Data from Gruber (2014, p. 252). Reproduced with kind permission of John Wiley & Sons.)

The analysis of variance is given below.

| Source | d.f. | Sum of Squares | Mean Square | $F$ |
|---|---|---|---|---|
| Brand | 3 | 84,663 | 28,221 | 271.214 |
| Error | 26 | 2706 | 104 | |
| Total | 29 | 87,369 | | |

We shall estimate the variance components and the variance of the variance components.

Using (62), we have that

$$\hat{\sigma}_e^2 = \frac{2706}{26} = 104 \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{84663 - 3(104)}{30 - (7^2 + 8^2 + 9^2 + 6^2)/30} = 3777.$$

Equation (62) found the expectations of the sum of squares. To find the expected mean square, this expectation is divided by the degrees of freedom. This could be done with the expectations calculated in Example 4 above. Using the random command in PROC GLM, we get the result below. This would be the same as in (62) if we divide by the degrees of freedom. Thus, for a particular data set, we can obtain the expected mean squares using SAS whether the data is balanced or not.

| Source | Type III Expected Mean Square |
|---|---|
| Brand | var(error) + 7.4444 var(brand) |

Using the formulae of Crump (1947) and Searle (1956), we obtain $k_1 = 0.07692, k_2 = -0.01033, k_3 = 0.01342, k_4 = 0.1791$, and $k_5 = 0.6768$. Substitution into (64) yields the estimates $v(\hat{\sigma}_e^2) = 772.543, \text{cov}(\hat{\sigma}_e^2, \hat{\sigma}_\alpha^2) = -103.749, v(\hat{\sigma}_\alpha^2) = 5.80007 \times 10^6$. □

***(iii) Calculation by Synthesis.*** The "synthesis" method of calculating numerical coefficients of $\hat{\sigma}^2$'s in expected values has been described in Section 2c(iv). We can also apply it to calculate squares and products of $\hat{\sigma}^2$'s in variances and covariances. We give the procedure for obtaining $E(T_A T_B)$ from which we can obtain $\text{cov}(T_A, T_B)$ using $E(T_A)$ and $E(T_B)$ based upon (42).

We first write $\mathbf{e} = \mathbf{X}_0\mathbf{b}_0$ with $\mathbf{X}_0 = \mathbf{I}$ and $\mathbf{b}_0 = \mathbf{e}$. The model (34) then becomes $\mathbf{y} = \mu\mathbf{1} + \sum_{\theta=0}^{K} \mathbf{X}_\theta \mathbf{b}_\theta$. Hartley (1967) then derives $E(T_A T_B)$ in the form

$$E(T_A T_B) = \sum_{\theta,\varphi=0}^{K} k\left(\sigma_\theta^2\sigma_\varphi^2, T_A T_B\right)\sigma_\theta^2\sigma_\varphi^2 + \sum_{\theta=0}^{K} h(\mu_{4,\theta}, T_A T_B)\mu_{4,\theta},$$

where by definition,

$$\mu_{4,0} = E(e_i^4) \quad \text{for} \quad i = 1, 2, \ldots N$$

and for $\theta = A, B, \ldots, K$

$$\mu_{4,\theta} = E(b_{\theta_j}^4) \quad \text{for} \quad \text{j} = 1, 2, \ldots, N_\theta.$$

With these definitions, the coefficients that Hartley (1967) gives are

$$\begin{aligned} h(\mu_{4\theta}, T_A T_B) &= \text{coefficient of } E(T_A T_B) \\ &= \sum_{j==1}^{N_\theta} T_A[x(\theta,j)] T_B[x(\theta,j)], \qquad (65) \\ k\left(\sigma_\theta^4, T_A T_B\right) &= \text{coefficient of } \sigma_\theta^4 \text{ in } E(T_A T_B) \\ &= \sum_{j=1}^{N_\theta} \sum_{j<j'} T_A[x(\theta,j) + x(\theta,j')] T_B[x(\theta,j) + x(\theta,j')] \\ &\quad - (N_\theta - 5) h(\mu_{4,\theta}, T_A T), \qquad (66) \end{aligned}$$

and

$$\begin{aligned} k\left(\sigma_\theta^2 \sigma_\varphi^2, T_A T_B\right) &= \text{coefficient of } \sigma_\theta^2 \sigma_\varphi^2 \text{ in } E(T_A T_B) \\ &= \frac{1}{2} \sum_{j=1}^{N_\theta} \sum_{j'=1}^{N_\varphi} T_A[x(\theta,j) + x(\varphi,j')] T_B[x(\theta,j) + x(\varphi,j')] \\ &\quad - N_\theta h(\mu_{4,\varphi}, T_A T_B) - N_\varphi h(\mu_{4,\theta}, T_A T_B). \qquad (67) \end{aligned}$$

Thus for $h(\mu_{4\theta}, T_A T_B)$, we use columns $\mathbf{X}_\theta$ as "data" vectors in $T_A$ and $T_B$. In $k(\sigma_\theta^4, T_A T_B)$, we add pairs of different columns of $\mathbf{X}_\theta$ and use the sums as "data" vectors in $T_A$ and $T_B$. These results are quite general and apply to any quadratic forms of the observations, including the use of $T_A$ in place of $T_B$ to obtain $E(T_A^2)$ and hence $v(T_A)$. Furthermore, the results are all in terms of variances and fourth moments. No particular form of the distribution has been assumed for the random variables. The formulae are well-suited computationally for obtaining coefficients, numerically in specific situations. However, with large amounts of data, the calculations would be extensive. We could also use these formulae to find coefficients algebraically. However, in most cases, the details involved would be most tedious. A simple example follows.

**Example 6 Illustration of the Results by Finding the Variance of $S^2$** Hartley (1967) illustrates his results by finding the variance of the usual unbiased estimator

of the sample variance

$$s^2 = s^2(x) = \frac{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{(n-1)}$$

as

$$v(s^2) = E(s^2 s^2) - \sigma^4,$$

where

$$E(s^2 s^2) = k_{00}\sigma^4 + h_0 \mu_{4,0}.$$

By (65),

$$\begin{aligned} h_0 &= \sum_{j=1}^{N_\theta} \{[x(0,j)]\}^2 \\ &= \sum_{j=1}^{n} [s^2(\text{column of } \mathbf{I}_n)]^2 \\ &= n\left[\frac{1 - n(1/n)^2}{n-1}\right]^2 = \frac{1}{n}. \end{aligned}$$

Furthermore, by (66),

$$\begin{aligned} k_{00} &= \sum_{j=1}^{N_\theta} \sum_{j'<j}^{N_\theta} \{s^2[x(0,j) + x(0,j')]\}^2 - (N_\theta - 5)h_0 \\ &= \sum_{j=1}^{n} \sum_{j<j'}^{n} [s^2(\text{sum of 2 columns of } \mathbf{I}_n)]^2 - \frac{n-5}{n} \\ &= \frac{(n-1)n}{2}\left[\frac{2 - n(2/n)^2}{(n-1)}\right]^2 - \frac{n-5}{n} = \frac{n^2 - 2n + 3}{n(n-1)}. \end{aligned}$$

Hence,

$$\begin{aligned} v(s^2) &= k_{00}\sigma^4 + h_0\mu_{4,0} - \sigma^4 \\ &= \left[\frac{n^2 - 2n + 3}{n(n-1)} - 1\right]\sigma^4 + \frac{\mu_{4,0}}{n} = \frac{3-n}{n(n-1)}\sigma^4 + \frac{\mu_{4,0}}{n}. \end{aligned}$$

The above result can also be obtained directly (see Exercise 4). With normality assumptions $\mu_{4,0} = 3\sigma^4$ and the result reduces to the familiar $v(s^2) = 2\sigma^4/(n-1)$. □

## 3. ADJUSTING FOR BIAS IN MIXED MODELS

We indicated in Section 2c(ii) that with unbalanced data, the analysis of variance method for mixed models leads to biased estimators of variance components. There is, of course, a dual problem with mixed models—estimation of both the fixed effects and the variance components of the random effects. Here, we confine attention to estimating, just the variance components. In some situations, this would be exactly what would be done in practice. For example, with genetic data effects that are often considered fixed, such as year effects might be of little interest compared to genetic variance components. On the other hand, if trends in the year effects were of interest, their estimation together with that of variance components would be considered simultaneously. We now consider the dual estimation problem.

Henderson (1953) presents the method. He first uses the data to estimate the fixed effects of the model. He then adjusts the data by these estimators. Then, he estimates the variance components using the adjusted data. The design of the whole procedure is such that the presence of fixed effects does not cause the variance components to be biased. The analysis of variance estimators would be biased. This method produces unbiased estimators. However, Searle (1968) shows that the method is not uniquely defined. Furthermore, we cannot use certain simplified forms of the method whenever the model includes interactions between the fixed effects and the random effects. We now consider these points. We follow Searle (1968) closely but do not repeat the details here.

### a. General Method

We consider the general model (34) in the form

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_f\mathbf{b}_f + \mathbf{X}_r\mathbf{b}_r + \mathbf{e}. \tag{68}$$

In the model (68), we represent all of the fixed effects other than $\mu$ by $\mathbf{b}_f$ and all of the random effects by $\mathbf{b}_r$. We take $E(\mathbf{b}_r) = \mathbf{0}$. Then, $E(\mathbf{b}_r\mathbf{b}_r') = \text{var}(\mathbf{b}_r)$. This is the variance–covariance matrix of the random effects. Suppose and estimator of the fixed effects $\mathbf{b}_f$ is $\tilde{\mathbf{b}}_f = \mathbf{L}\mathbf{y}$. Then, $\mathbf{z} = \mathbf{y} - \mathbf{X}_f\tilde{\mathbf{b}}_f$ is a vector of data adjusted by the vector $\mathbf{b}_f$. Substitution from (68) shows that the model for $\mathbf{z}$ contains no terms on $\mathbf{b}_f$ provided that $\mathbf{L}$ is a generalized inverse of $\mathbf{X}_f$. Under this condition, the analysis of variance method applied to $\mathbf{z}$ will yield unbiased estimators of the variance components. However, the fact that $\mathbf{L}$ has only to be a generalized inverse of $\mathbf{X}_f$ indicates the lack of uniqueness in the method.

### b. A Simplification

The calculations involved in applying the analysis of variance method to $\mathbf{y}$, particularly those involving the random effects $\mathbf{X}_r\mathbf{b}_r$ have been documented in the preceding section. In $\mathbf{z} = \mathbf{y} - \mathbf{X}_r\tilde{\mathbf{b}}_f$, the term in the random effects is $\mathbf{X}_r - \mathbf{X}_f\mathbf{L}\mathbf{X}_r$.

For the case where $\mathbf{X}_f\mathbf{L}\mathbf{X}_r = \mathbf{0}$, application of the analysis of variance method to $\mathbf{z}$ would, so far as random effects are concerned, be the same as applying the method to $\mathbf{y}$. To be more specific, suppose we chose $\mathbf{L}$ such that the model for $\mathbf{z}$ is

$$\mathbf{z} = \mu^*\mathbf{1} + \mathbf{X}_r\mathbf{b}_r + \mathbf{Z}\mathbf{e}, \tag{69}$$

for $\mu^*$ being a scalar (not necessarily equal to $\mu$) and for $\mathbf{Z}$ being the same matrix. The analysis of variance method applied to (69) would involve no fixed effects. Although treatment of the error terms in (69) would differ from that of the error terms in (68), treatment of the random effects would be the same as when using (68). Therefore, apart from calculations relating to $\sigma_e^2$, using the analysis of variance method on (69) would be the same as using it on (68) with the fixed effects ignored. To achieve this, Searle (1968) shows that $\mathbf{L}$ need not be a generalized inverse of $\mathbf{X}_f$ but has to satisfy three conditions. They are:

$$\mathbf{X}_f\mathbf{L}\mathbf{X}_f = \mathbf{0}, \tag{70}$$

$$\mathbf{X}_f\mathbf{L} \text{ has its row sums equal,} \tag{71}$$

and

$$\mathbf{X}_f - \mathbf{X}_f\mathbf{L}\mathbf{X}_f \text{ has all its rows the same.} \tag{72}$$

Although the non-unique condition on $\mathbf{L}$ that $\mathbf{X}_f\mathbf{L}\mathbf{X}_f = \mathbf{X}_f$ has been replaced by the conditions in (70)–(72), they do not necessarily determine $\mathbf{L}$ uniquely. Furthermore, these conditions imply that the model for $\mathbf{y}$ must not contain interactions between fixed and random effects. This is a severe limitation on the method.

### c. A Special Case: Henderson's Method 2

Method 2, described by Henderson (1953) is simply one specific way that we can carry out the simpler form of the generalized method. Henderson's method 2 estimates $\mathbf{b}_f$ as $\tilde{\mathbf{b}}_f = \mathbf{L}\mathbf{y}$. It uses an $\mathbf{L}$ that satisfies (70), (71), and (72), and then uses the analysis of variance method on $\mathbf{y} - \mathbf{X}_f\tilde{\mathbf{b}}_f$. Even though Henderson's method 2 is just one way of executing the simpler form of the generalized model, it suffers from the limitation already alluded to. It cannot be used whenever the model has interactions between fixed and random effects. Although Henderson (1953) does not state this explicitly, his example does not have interactions between fixed and random effects. There, the fixed effects in a study of dairy production are years. The random effects are herds, sires, and herds-by-sires interactions. There are no interactions of years with herds and/or sires.

To use Henderson's method 2, we first estimate $\mathbf{b}_f$ by least squares assuming temporarily, and for this purpose only, that $\mu = 0$ and that the random effects are fixed. This leads to the equations

$$\mathbf{X}'\mathbf{X}\begin{bmatrix}\tilde{\mathbf{b}}_f\\ \tilde{\mathbf{b}}_r\end{bmatrix} = \begin{bmatrix}\mathbf{X}_f'\mathbf{X}_f & \mathbf{X}_f'\mathbf{X}_r\\ \mathbf{X}_r'\mathbf{X}_f & \mathbf{X}_r'\mathbf{X}_r\end{bmatrix}\begin{bmatrix}\mathbf{b}_f\\ \mathbf{b}_r\end{bmatrix} = \begin{bmatrix}\mathbf{X}_f'\mathbf{y}\\ \mathbf{X}_r'\mathbf{y}\end{bmatrix}. \tag{73}$$

It is the manner in which (73) is solved that leads to the solution $\tilde{\mathbf{b}}_f$ being $\tilde{\mathbf{b}}_f = \mathbf{Ly}$ satisfying (70), (71), and (72). The essential part of the solution is picking a generalized inverse of $\mathbf{X}'\mathbf{X}$ in the manner described in Section 1 of Chapter 1. Recall that in finding a generalized inverse of $\mathbf{X}'\mathbf{X}$ we strike out rows and columns to reduce it to full rank. This should be done in such a way so that as many as possible of these rows and columns are through $\mathbf{X}_f'\mathbf{X}_f$. Searle (1968) gives details of this process. Despite being able to specify the method in this manner, the method suffers from the deficiencies already alluded to. It is not uniquely specified. It cannot be used in the presence of interactions between fixed and random effects. For these reasons, we do not recommend its use.

## 4. FITTING CONSTANTS METHOD (HENDERSON'S METHOD 3)

Fitting the linear models of Chapter 5–8 is often referred to as the technique of fitting constants, as mentioned in Chapter 4, because the effects of fixed-effects models are sometimes called constants. We now describe a third method of fitting variance components that is based on fitting fixed-effects models. Accordingly, it is called the *fitting constants method,* or Henderson's method 3, after Henderson (1953). For whatever model being used, the method uses reductions in the sums of squares due to fitting this model and the different sub-models thereof, in the manner of Chapter 6, 7, and 8. These reductions, the $R(\ )$-terms of those chapters, are used in the analysis of variance method in the same manner as are the SS-terms, the analogous sums of squares of the analysis of variance method. One estimates the variance components by equating each computed reduction to its expected value—its expected value under the full model. We describe the general properties of the method and then illustrate its application in the two-way classification. The presentation follows closely that of Searle (1968).

### a. General Properties

We rewrite the general model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ as

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}. \tag{74}$$

The portioning simply divides $\mathbf{b}$ into two groups with no thought for whether the groups represent fixed or random effects. We will denote the reduction in sum of squares due to fitting this model by $R(\mathbf{b}_1, \mathbf{b}_2)$. For the moment, we are concerned with finding the expected values of $R(\mathbf{b}_1, \mathbf{b}_2)$ and of the reduction in the sum of squares due to fitting the sub-model

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{e}. \tag{75}$$

Both expectations will be taken under the full model (74).

Denoting the reduction in the sum of squares due to fitting (75) by $R(\mathbf{b}_1)$, we write

$$R(\mathbf{b}_2|\mathbf{b}_1) = R(\mathbf{b}_1, \mathbf{b}_2) - R(\mathbf{b}_1) \tag{76}$$

in the manner of Section 3a of Chapter 6. We will show that the expected value of (76) under the model (74) involves only $\sigma_e^2$ and

$$E(\mathbf{b}_2\mathbf{b}_2') = \text{var}(\mathbf{b}_2) + E(\mathbf{b}_2)E(\mathbf{b}_2)', \tag{77}$$

and it does not involve $\mathbf{b}_1$. Consequently, the fitting constants method, by judicious choice of sub-models represented by $\mathbf{b}_1$ in (76), yields unbiased estimators of the variance components of the full model. These estimators are uncomplicated by any fixed effects that may be in the model.

First, we slightly modify equation (2) for $E(\mathbf{y}'\mathbf{Q}\mathbf{y})$. In the general model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, the vector $\mathbf{b}$ can be fixed, random, or mixed. Adopting the convention that for a fixed effect $E(b_i) = b_i$ enables $E(\mathbf{b})$ to be defined whatever be the nature of $\mathbf{b}$. Thus from (2),

$$\begin{aligned} E(\mathbf{y}'\mathbf{Q}\mathbf{y}) &= \text{tr}[\mathbf{Q}\{\mathbf{X}\text{Var}(\mathbf{b})\mathbf{X}' + \sigma_e^2\mathbf{I}\}] + E(\mathbf{b}')\mathbf{X}'\mathbf{Q}\mathbf{X}E(\mathbf{b}) \\ &= \text{tr}[\mathbf{X}'\mathbf{Q}\mathbf{X}E(\mathbf{b}\mathbf{b}')] + \sigma_e^2\text{tr}(\mathbf{Q}). \end{aligned}$$

In this form, $E(\mathbf{y}'\mathbf{Q}\mathbf{y})$ is suitable for considering the models (74) and (75).

In fitting (74), the reduction sum of squares is the same as that of equation (14) of Chapter 5. Thus,

$$R(\mathbf{b}_1, \mathbf{b}_2) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}, \tag{78}$$

where $(\mathbf{X}'\mathbf{X})^-$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Taking the expectation of (78) gives

$$\begin{aligned} ER(\mathbf{b}_1, \mathbf{b}_2) &= \text{tr}\{((\mathbf{X}'\mathbf{X})E(\mathbf{b}\mathbf{b}')\} + \sigma_e^2 r(\mathbf{X}) \\ &= \text{tr}\left\{ \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} E(\mathbf{b}\mathbf{b}') \right\} + \sigma_e^2 r(\mathbf{X}). \end{aligned}$$

Similarly, when fitting (75), the reduction in the sum of squares is

$$R(\mathbf{b}_1) = \mathbf{y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{y},$$

with

$$\begin{aligned} ER(\mathbf{b}_1) &= \text{tr}\{\mathbf{X}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{X}E(\mathbf{b}\mathbf{b}')\} + \sigma_e^2 r(\mathbf{X}_1) \\ &= \text{tr}\left\{ \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 \\ \mathbf{X}_2'\mathbf{X}_1 \end{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^- \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \end{bmatrix} E(\mathbf{b}\mathbf{b}') \right\} + \sigma_e^2 r(\mathbf{X}_1) \\ &= \text{tr}\left\{ \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{X}_2 \end{bmatrix} E(\mathbf{b}\mathbf{b}') \right\} + \sigma_e^2 r(\mathbf{X}_1). \end{aligned}$$

Hence, the expected value of $R(\mathbf{b}_2|\mathbf{b}_1)$ is

$$\begin{aligned} E[R(\mathbf{b}_2|\mathbf{b}_1)] &= E[R(\mathbf{b}_2, \mathbf{b}_1) - R(\mathbf{b}_1)] \\ &= \text{tr}\{\mathbf{X}_2'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1']\mathbf{X}_2 E(\mathbf{b}_2\mathbf{b}_2')\} + \sigma_e^2[r(\mathbf{X}) - r(\mathbf{X}_1)]. \end{aligned} \tag{79}$$

As forecast, the only $\mathbf{b}$-term involved here is $\mathbf{b}_2$. The expectation of $R(\mathbf{b}_2|\mathbf{b}_1)$ is a function simply of $E(\mathbf{b}_2\mathbf{b}_2')$ and $\sigma_e^2$. It does not involve $E(\mathbf{b}_1\mathbf{b}_1')$ or $E(\mathbf{b}_1\mathbf{b}_2')$. Observe that this result has been derived without any assumptions on the form of $E(\mathbf{b}\mathbf{b}')$.

The result in (79) has important consequences. It means that if the $\mathbf{b}$-vector of one's model can be partitioned into two parts $\mathbf{b}_1$ and $\mathbf{b}_2$, where $\mathbf{b}_2$ contains just random effects, then $ER(\mathbf{b}_2|\mathbf{b}_1)$ as given in (79) contains only $\sigma_e^2$ and the variance components relating to those random effects. Thus, when $\mathbf{b}_1$ represents all the fixed effects, $ER(\mathbf{b}_2|\mathbf{b}_1)$ contains no terms due to those fixed effects. This is the value of the method of fitting constants to the mixed model. The method yields estimates of the variance components unaffected by the fixed effects. Furthermore, in the random model, when $\mathbf{b}_1$ contains random effects, $ER(\mathbf{b}_2|\mathbf{b}_1)$ contains no terms arising from $\text{var}(\mathbf{b}_1)$. More importantly, it does not contain any terms arising from any covariance between the elements of $\mathbf{b}_1$ and $\mathbf{b}_2$. Hence, even if the model is such that the terms in $\mathbf{b}_1$ are correlated with terms in $\mathbf{b}_2$, the expectation in (79) does not involve this correlation. It depends solely on the second moments in $\mathbf{b}_2$ (and on $\sigma_e^2$).

Compared with the analysis of variance method, the immediate importance of the fitting constants method lies in its appropriateness for the mixed model for which it yields variance component estimators that are unbiased for the fixed effects. Therefore, it is the preferred method for mixed models. Its disadvantage is that it involves calculating generalized inverses that will be very large in models having large number of effects in them. This difficulty can arise in calculating both reductions in sums of squares and the coefficients in the $\sigma^2$'s in their expectations. Hartley's (1967) method of synthesis described in Section 2d(iii) can be used as one means of calculation. Other available shortcuts can be found in Gaylor et al. (1970).

We now consider the application of this method to the two-way classification.

### b. The Two-Way Classification

Equation (12) is that of the two-way classification model. Table 7.8 gives the reduction in the sums of squares for fitting the fixed-effects version of this model. This table includes

$$\begin{aligned} R(\alpha|\mu) &= R(\mu, \alpha) - R(\mu) \\ R(\beta|\mu, \alpha) &= R(\mu, \alpha, \beta) - R(\mu, \alpha) \\ R(\gamma|\mu, \alpha, \beta) &= R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta) \end{aligned} \tag{80}$$

and

$$\text{SSE} = \sum y^2 - R(\mu, \alpha, \beta, \gamma).$$

We shall use these terms in the fitting constants method of estimating variance components. To do so, we need their expected values.

***(i) Expected Values.*** As usual, the expected value of SSE in (80) is $(N-s)\sigma_e^2$. Taking $R(\gamma|\mu,\alpha,\beta)$ next, its expected value can be derived from (79). However, we cannot obtain the expected values of $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$ directly from (79). This is because (79) is the expected value of $R(\mathbf{b}_2|\mathbf{b}_1) = R(\mathbf{b}_1,\mathbf{b}_2) - R(\mathbf{b}_1)$. This is the difference between two $R(\cdot)$-terms. One of them is for the full model; the other is for a sub-model. This is the only kind of term to which (79) applies. An example is $R(\gamma|\mu,\alpha,\beta)$ of (80). In contrast, (79) does not apply to $R(\cdot|\cdot)$-terms that are differences between two $R(\cdot)$-terms that are both for sub-models. For this reason, with the full model involving $\mu,\alpha,\beta$, and $\gamma$, (79) does not apply to $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$ of (80).

Although (79) cannot be used directly on $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$, it can be utilized by considering certain sums of squares of the terms in (80) that involve $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$. For example, (79) applies to

$$R(\alpha,\beta,\gamma|\mu) = R(\mu,\alpha,\beta,\gamma) - R(\mu). \tag{81}$$

This is the sum of the first three terms in (80):

$$R(\alpha,\beta,\gamma|\mu) = R(\alpha|\mu) + R(\beta|\mu,\alpha) + R(\gamma|\mu,\alpha,\beta).$$

Similarly, (79) applies to

$$R(\beta,\gamma|\mu,\alpha) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha) \tag{82}$$

which is

$$R(\beta,\gamma|\mu,\alpha) = R(\beta|\mu,\alpha) + R(\gamma|\mu,\alpha,\beta).$$

Equating observed values of $R(\cdot|\cdot)$-terms to their expected values to obtain variance component estimators, using (81) and (82) in place of $R(\alpha|\mu)$ and $R(\beta|\alpha,\mu)$ of (80) yields equations that are linear combinations of those that would arise using (80). Therefore, the estimators will be the same. Table 10.1 shows the form taken by the

**TABLE 10.1 Reductions in Sum of Squares for Estimating Variance Components in a Two-Way Classification Interaction, Random Model, Unbalanced Data**

| Reduction in Sum of Squares | Computing Formula[a] | Expected Values[b] |
|---|---|---|
| $R(\alpha,\beta,\gamma\|\mu) = R(\mu,\alpha,\beta,\gamma) - R(\mu)$ | $= T_{AB} - T_\mu$ | $h_1\sigma_\alpha^2 + h_2\sigma_\beta^2 + h_3\sigma_\gamma^2 + (s-1)\sigma_e^2$ |
| $R(\beta,\gamma\|\mu,\alpha) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha)$ | $= T_{AB} - T_A$ | $h_4\sigma_\beta^2 + h_5\sigma_\gamma^2 + (s-a)\sigma_e^2$ |
| $R(\gamma\|\mu,\alpha,\beta) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta)$ | $= T_{AB} - R(\mu,\alpha,\beta)$ | $h_6\sigma_\gamma^2 + s^*\sigma_e^2$ |
| SSE $= \sum y^2 - R(\mu,\alpha,\beta,\gamma)$ | $= T_0 - T_{AB}$ | $(N-s)\sigma_e^2$ |

[a]The $T$'s are defined in (19) and $R(\mu,\alpha,\beta)$ is defined in (63) of Section 2d(i) of Chapter 7.
[b]The $h$'s come from (79) and are given in Section 4e of Chapter 11.
$s^* = s - a - b + 1$.

expected values of this reduction (81), (82), and the last two terms of (80). The table also shows the computing formulas for the reductions. We discuss these and the $h$-coefficients of the $\sigma^2$'s (which would be derived from (79) later). The coefficients of the $\sigma_e^2$'s have already been obtained from (79).

**(*ii*) *Estimation.*** The nature of (79) and of the reductions shown in Table 10.1 ensures that the expectations of those reductions involve successively more variance components, one at a time reading from the bottom up.

Estimation of these components from Table 10.1 is therefore quite straightforward. The estimators are

$$\begin{aligned}
\hat{\sigma}_e^2 &= \frac{\text{SSE}}{(N-s)} \\
\hat{\sigma}_\gamma^2 &= \frac{\left[R(\gamma|\mu,\alpha,\beta) - (s-a-b+1)\hat{\sigma}_e^2\right]}{h_6} \\
\hat{\sigma}_\beta^2 &= \frac{\left[R(\beta,\gamma|\mu,\alpha) - h_5\hat{\sigma}_\gamma^2 - (s-a)\hat{\sigma}_e^2\right]}{h_4}
\end{aligned} \tag{83}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{\left[R(\alpha,\beta,\gamma|\mu) - h_2\hat{\sigma}_\beta^2 - h_3\hat{\sigma}_\gamma^2 - (s-1)\hat{\sigma}_e^2\right]}{h_1}$$

Once we obtain the $R$'s and the $h$'s, we can calculate these estimators easily. We turn to this now.

**(*iii*) *Calculation.*** Expressions for calculating the $R(\cdot)$-terms of Table 10.1 are given in equations (58)–(63) of Section 2d(i) of Chapter 7. Most of them are the same as the $T$'s given in (19) of this chapter; that is,

$$\begin{aligned}
R(\mu) &= T_\mu, & R(\mu,\alpha) &= T_A, \\
R(\mu,\alpha,\beta,\gamma) &= T_{AB} \quad \text{and} & \textstyle\sum y^2 &= T_0.
\end{aligned} \tag{84}$$

These are easy to calculate, as in (19), and lead to the computing formulae shown in Table 10.1. Notice that the only term that is not part of the analysis of variance method is $R(\mu,\alpha,\beta)$. Calculation of this is given in equations (63)–(65) in Section 2d(i) of Chapter 7 and repeated again in Chapter 11 (see the web page). For the moment, we concern ourselves with the general methodology rather than the specific applications. Therefore, details of calculating $R(\mu,\alpha,\beta)$ and the $h$'s of Table 10.1 are left until Chapter 11 (the web page).

In passing, we may note that because the reductions in Table 10.1 are largely functions of the $T$'s, most of the $h$'s are correspondingly functions of coefficients of $\sigma^2$'s in expected values of T's. Equations (38) contain a general expression for these coefficients. Chapter 11 (the web page) shows full details.

**TABLE 10.2 An Alternative Set of Reductions in Sum of Squares for Estimating Variance Components in a Two-Way Classification Interaction, Random Model, Unbalanced Data**

| Reduction in Sum of Squares | Computing Formula[a] | Expected Values[b] |
|---|---|---|
| $R(\alpha,\beta,\gamma\|\mu) = R(\mu,\alpha,\beta,\gamma) - R(\mu)$ | $= T_{AB} - T_{\mu}$ | $h_1\sigma_{\alpha}^2 + h_2\sigma_{\beta}^2 + h_3\sigma_{\gamma}^2 + (s-1)\sigma_e^2$ |
| $R(\alpha,\gamma\|\mu,\beta) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\beta)$ | $= T_{AB} - T_B$ | $h_7\sigma_{\alpha}^2 + h_8\sigma_{\gamma}^2 + (s-B)\sigma_e^2$ |
| $R(\gamma\|\mu,\alpha,\beta) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta)$ | $= T_{AB} - R(\mu,\alpha,\beta)$ | $h_6\sigma_{\gamma}^2 + s^*\sigma_e^2$ |
| $\text{SSE} = \sum y^2 - R(\mu,\alpha,\beta,\gamma)$ | $= T_0 - T_{AB}$ | $(N-s)\sigma_e^2$. |

[a]The $T$'s are defined in (19) and $R(\mu,\alpha,\beta)$ is defined in (63) of Section 2d(i) of Chapter 7.
[b]The $h$'s come from (79) and are given in Section 4e of Chapter 11.
$s^* = s - a - b + 1$.

### c. Too Many Equations

Table 10.1 contains no term $R(\mu,\beta) = T_B$. This is because the table is based on the reductions in the sum of squares shown in (80). These in turn come from the first part of Table 7.8. This deals, in the fixed-effect model, with the fitting of $\alpha$ before $\beta$. In this context, $R(\mu,\beta)$ does not arise. On the other hand, $R(\mu,\beta) = R(\mu) + R(\beta|\mu)$, comes from the second part of Table 7.8, concerned with fitting $\beta$ before $\alpha$. However, observe that there is nothing sacrosanct about either part of the table as far as estimation of variance components in the model is concerned. In (80), we used the first part. However, we could have just as well-used the second. Rearrangement of the reductions in sums of squares therein, in the manner of Table 10.1, yields Table 10.2. Table 10.2 is exactly the same as Table 10.1 except for the second entry that involves $R(\mu,\beta)$ instead of $R(\mu,\alpha)$.

Equating the reductions to their expected values yields the following estimators of the variance components:

$$\begin{aligned}
\hat{\sigma}_e^2 &= \frac{\text{SSE}}{(N-s)} \\
\hat{\sigma}_{\gamma}^2 &= \frac{\left[R(\gamma|\mu,\alpha,\beta) - (s-a-b+1)\hat{\sigma}_e^2\right]}{h_6} \\
\hat{\sigma}_{\alpha}^2 &= \frac{\left[R(\alpha,\gamma|\mu,\beta) - h_8\hat{\sigma}_{\gamma}^2 - (s-b)\hat{\sigma}_e^2\right]}{h_7}
\end{aligned} \tag{85}$$

and

$$\hat{\sigma}_{\beta}^2 = \frac{\left[R(\alpha,\beta,\gamma|\mu) - h_1\hat{\sigma}_{\beta}^2 - h_3\hat{\sigma}_{\gamma}^2 - (s-1)\hat{\sigma}_e^2\right]}{h_2}.$$

The estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_{\gamma}^2$ in (85) are the same as those in (83) but $\hat{\sigma}_{\alpha}^2$ and are not.

The question immediately arises as to which estimators should be used, (83) or (85)? Unfortunately, there is no satisfactory answer to this question. Indeed, there is almost no answer at all. In the fixed-effects model, there is often good reason for choosing between fitting $\beta$ after $\alpha$ or $\alpha$ after $\beta$. However, for the random-effects model, there appears to be no criteria for making this choice when using the reductions in the sums of squares to estimate variance components. It means, in effect, that we can have, in the fitting constants method, more equations than variance components. For example, Tables 10.1 and 10.2 provide between them five equations in four variance components.

An unsolved difficulty with the fitting constants method of estimation is that it can yield more equations than there are components to be estimated. Moreover, it provides no guidance as to which equations should be used. The difficulty can assume some magnitude in multi-classification models. For example, there are six sets in a three-way classification model (see Table 8.2). Not only can each of these sets be used on their own. Combinations of terms from them can also be used. For example, in Tables 10.1 and 10.2, the last two lines are the same. These two lines and the second line of each table could be used to provide estimators. This is the principle of procedures considered by Harville (1967) and Low (1964).

A criterion that could have some appeal for deciding on which reductions to use is that they should add up to the total sum of squares corrected for the mean, $\mathrm{SST}_m = \sum y^2 - T_\mu$. Although the reductions listed in Tables 10.1 and 10.2 do not meet this requirement explicitly, they are linear combinations of reductions that do so and therefore provide the same estimators. For example, the terms in Table 10.1 are linear combinations of those in (80) which do add up to $\mathrm{SST}_\mathrm{m}$. One feature of this criterion is that the resulting estimators come from reductions that account for the total observed variability in the $y$'s, and they are reductions with known properties in fixed-effects models. This criterion would confine us to using sets of reductions like those in Tables 10.1 and 10.2 and would preclude using combinations of terms from these tables. On the other hand, using combinations is attractive, because, for example, Table 10.1 excludes $R(\mu, \beta)$ and Table 10.2 excludes $R(\mu, \alpha)$. Intuitively, one might feel these terms should not be omitted.

Knowing, as we do, certain properties in the analysis of variance with balanced data suggests that whatever reductions are used for estimating variance components from unbalanced data, they should be such as to reduce the resulting estimators to the analysis of variance estimators when the data *are* balanced, that is, when the $n_{ij}$'s are equal. For example, (80) reduces to Table 7.9 when $n_{ij} = n$ for all $i$ and $j$.

One possible way of overcoming the situation of having more equations than variance components is to apply "least squares" as suggested by Robson (1957). Arraying all calculated reductions as a vector $\mathbf{r}$ let us suppose that $E(\mathbf{r}) = \mathbf{A}\sigma^2$. Then, $\mathbf{r} = \mathbf{A}\hat{\sigma}^2$ are the equations we would like to solve for $\hat{\sigma}^2$. However, when there are more equations than variance components these equations will usually not be consistent.[2] Nevertheless, provided the reductions in $\mathbf{r}$ are linearly independent and $\mathbf{A}$ has full-column rank, we could estimate $\hat{\sigma}^2$ by "least squares" as $\hat{\sigma}^2 = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{r}$.

[2] S. R. Searle thanks D. A. Harville for bringing this to his attention.

### d. Mixed Models

The fitting constants method of estimation applies equally as well to mixed models as to random models. Indeed, for mixed models, it provides unbiased estimators. The analysis of variance method does not. As has already been explained, this arises from (79). Based on that result, we use only those reductions that have no fixed effects in their expected values. For example, in the two-way classification model with $\alpha$'s as fixed effects, we would use the last three lines of Table 10.1. By (79), they will have no fixed effects in their expectations. Furthermore, they provide unbiased estimators of $\sigma_e^2$, $\sigma_\gamma^2$, and $\sigma_\beta^2$. The one entry in Table 10.2 that differs from Table 10.1 is $R(\alpha, \gamma, \mu, \beta)$. It has, by (79), an expected value that is not free of the fixed $\alpha$-effects and so cannot be used. Therefore, the basis of estimation in the two-way classification model having $\alpha$'s as fixed effects is the last three lines of Table 10.1.

The principles we include here are quite straightforward and extend readily to multi-classification mixed models.

We can mention here variations on Henderson's method 2 of adjusting for bias in mixed models. As shown in equation (73), Henderson's method 2 temporarily assumes the random effects are fixed in order to solve the normal equations for the fixed effects.

An alternative is to temporarily ignore the random effects, and solve normal equations for fixed effects as $\tilde{\mathbf{b}}_f = (\mathbf{X}_f'\mathbf{X}_f)^-\mathbf{X}_f'\mathbf{y}$. Using (68) as the model for $\mathbf{y}$, we then adjust the data to be

$$\mathbf{z} = \mathbf{y} - \mathbf{X}_f\tilde{\mathbf{b}}_f = [\mathbf{I} - \mathbf{X}_f(\mathbf{X}_f'\mathbf{X}_f)^-\mathbf{X}_f']\mathbf{y} = [\mathbf{I} - \mathbf{X}_f(\mathbf{X}_f'\mathbf{X}_f)^-\mathbf{X}_f'](\mathbf{X}_r\mathbf{b}_r + \mathbf{e}).$$

Two possibilities are available with **z:** the analysis of variance method and the fitting constants method. Zelen (1968) suggested that the fitting constants method for $\mathbf{z}$ is equivalent to the use of the fitting constants method directly on y. Searle (1969) demonstrates the details of this.

### e. Sampling Variances of Estimators

Each $R(\cdot)$ reduction used in the fitting constants method can be expressed in the form $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ for some matrix $\mathbf{X}$. Therefore, on the basis of the normality assumptions for both the error terms and the random effects, we can derive the sampling variance of each reduction using Theorem 5 of Chapter 2. In a similar manner, we can derive covariances between reductions. We have that when $\mathbf{y} \sim N(\mu, \mathbf{V})$,

$$\text{cov}(\mathbf{y}'\mathbf{P}\mathbf{y}, \mathbf{y}'\mathbf{Q}\mathbf{y}) = 2\text{tr}(\mathbf{PVQV}) + 4\mu'\mathbf{PVQ}\mu.$$

In this way, we can develop sampling variances of variance components, since the estimators are linear combinations of these reductions. The details are somewhat lengthy, involving extensive matrix manipulations. Rohde and Tallis (1969) give

general results applicable to both variance and covariance. Low (1964) and Harville (1969c) discuss specific cases.

Al Sarraj and von Rosen (2008) using ideas of Kelly and Mathew (1994) present a method of perturbing the variance component estimators obtained by Henderson's method 3 so that they have smaller mean square error.

## 5. ANALYSIS OF MEANS METHODS

Data in which every subclass of the model contains observations can, in fixed-effects models, be analyzed in terms of the means of the sub-most subclasses. Two such analyses are discussed in Section 3c of Chapter 8. We can use the mean squares of these analyses for estimating variance components in random and mixed models. Table 10.3 shows expected values of these mean squares for the random model. We

**TABLE 10.3 Expected Values of Mean Squares in Two Analyses of Means of The Two-Way Classification Interaction Random Model Having All $N_{ij} > 0$**

a. Unweighted Means Analysis (Table 8.12)[a]

$$
\begin{aligned}
E(\mathrm{MSA}_w) &= b\sigma_\alpha^2 + \sigma_\gamma^2 + n_h\sigma_e^2 \\
E(\mathrm{MSB}_w) &= a\sigma_\beta^2 + \sigma_\gamma^2 + n_h\sigma_e^2 \\
E(\mathrm{MSAB}_w) &= \sigma_\gamma^2 + n_h\sigma_e^2 \\
E(\mathrm{MSE}) &= \sigma_e^2
\end{aligned}
$$

b. Weighted Means Analysis (Table 8.18)[b]

$$
E(\mathrm{MSA}_w) = \frac{1}{(a-1)b}\left(\sum_{i=1}^{a} w_i - \frac{\sum_{i=1}^{a} w_i^2}{\sum_{i=1}^{a} w_i}\right)\left(b\sigma_\alpha^2 + \sigma_\gamma^2\right) + \sigma_e^2
$$

$$
E(\mathrm{MSA}_w) = \frac{1}{a(b-1)}\left(\sum_{j=1}^{b} v_j - \frac{\sum_{j=1}^{b} v_j^2}{\sum_{j=1}^{b} v_j}\right)\left(a\sigma_\beta^2 + \sigma_\gamma^2\right) + \sigma_e^2
$$

$$
\begin{aligned}
E(\mathrm{MSAB}) &= \sigma_\gamma^2 + n_h\sigma_e^2 \\
E(\mathrm{MSE}) &= \sigma_e^2
\end{aligned}
$$

[a] $n_h = \dfrac{\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}^{-1}}{ab}$.

[b] $w_i = \dfrac{b^2}{\sum_{j=1}^{b} n_{ij}^{-1}}$; $\quad v_j = \dfrac{a^2}{\sum_{i=1}^{a} n_{ij}^{-1}}$.

obtain estimators of the variance components in the usual manner of equating the mean squares to their expected values. The estimators are unbiased. Even though they are quadratic forms we can, under normality assumptions, obtain their variances using Theorem 4 of Chapter 2. We could also derive the variances by the method of "synthesis" described in Section 2d(iii). For mixed models, we shall only use the mean squares whose expectations contain no fixed effects to estimate the variance components. For example, if the $\alpha$'s are fixed effects in the two-way classification we will not use $\text{MSA}_u$ or $\text{MSA}_w$ of Table 10.3.

Extensions of Table 10.3 to multi-way classifications depends on the extension of Tables 8.12 and 8.18. This is particularly straightforward for the unweighted means analysis of Table 8.12. However, the need for having data in every subclass of the model still remains. Analysis of means cannot be made otherwise.

## 6. SYMMETRIC SUMS METHODS

Koch (1967a, 1968) suggests a method of estimating variance components based on symmetric sums of the observations, rather than the sum of squares. The method uses the fact that expected values of products of observations are linear combinations of the variance components. Therefore, sums of these products (and hence means of them) provide unbiased estimators of the components. We illustrate them in terms of the one-way classification.

Consider the random model for the one-way classification $y_{ij} = \mu + \alpha_i + e_{ij}$, where $E(\alpha_i) = E(e_{ij}) = 0, E(\alpha_i^2) = \sigma_\alpha^2$, and $E(e_{ij}^2) = \sigma_e^2$ for all $i$ and $j$, and all covariances are zero. Then, expected values of products of observations are as follows:

$$\begin{aligned} E(y_{ij}y_{i'j'}) &= \mu^2 + \sigma_\alpha^2 + \sigma_e^2 \quad \text{when} \quad i = i' \text{ and } j = j'; \\ &= \mu^2 + \sigma_\alpha^2 \quad \text{when} \quad i = i' \text{ and } j \neq j'; \\ &= \mu^2 \quad \text{when} \quad i \neq i'. \end{aligned} \tag{86}$$

We derive the estimators from the means of the different products in (86). We have that

$$\hat{\mu}^2 + \hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2}{N}, \tag{87}$$

$$\hat{\mu}^2 + \hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}\sum_{j'\neq j}^{n_i'} y_{ij}y_{ij'}}{\sum_{i=1}^{a} n_i(n_i - 1)} = \frac{\left(\sum_{i=1}^{a} y_{i.}^2 - \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2\right)}{(S_2 - N)} \tag{88}$$

where $S_2 = \sum_{i=1}^{a} n_i^2$, and

$$\hat{\mu}^2 = \frac{\sum_{i=1}^{a}\sum_{i'\neq i}^{a}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}} y_{ij}y_{i'j'}}{\sum_{i=1}^{a}\sum_{i'\neq i}^{a} n_i n_{i'}} = \frac{\left(y_{\cdot\cdot}^2 - \sum_{i=1}^{a} y_{i\cdot}^2\right)}{(N^2 - S_2)}. \tag{89}$$

We can obtain the estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_\alpha^2$ from these expressions.

These estimators are unbiased and consistent. Furthermore, they are identical to the analysis of variance estimators in the case of balanced data. However, as noted by Koch (1968), the variances of these estimators are functions of $\mu$. We can see evidence of this in $\hat{\sigma}_e^2$. Use (87) and (88) to write

$$\hat{\sigma}_e^2 = \mathbf{y}'\left(k_1\mathbf{I}_N - k_2\sum_{i=1}^{a} {}^{+}\mathbf{J}_{n_i}\right)\mathbf{y}, \tag{90}$$

where

$$k_1 = S_2/N(S_2 - N) \quad \text{and} \quad k_2 = 1/(S_2 - N). \tag{91}$$

When deriving the variance of (90) from Theorem 4 of Chapter 2, we will find the term in $\mu^2$ is

$$\begin{aligned} &4\mu^2\mathbf{1}'\left(k_1\mathbf{I}_N - k_2\sum_{i=1}^{a} {}^{+}\mathbf{J}_{n_i}\right)\left(\sigma_e^2\mathbf{I}_N + \sum_{i=1}^{a} {}^{+}\mathbf{J}_{n_i}\right)\left(k_1\mathbf{I}_N - k_2\sum_{i=1}^{a} {}^{+}\mathbf{J}_{n_i}\right)\mathbf{1} \\ &\quad = 4\mu^2\sum_{i=1}^{a}\left(\sigma_e^2 + n_i\sigma_\alpha^2\right)(k_1 - n_i k_2)^2. \end{aligned}$$

The above expression is non-zero for unequal $n_i$. It is zero when the $n_i$ are equal. Hence, for unbalanced data, the variance of $\hat{\sigma}_e^2$ derived from (87) and (88) is a function of $\mu$. We could also show this for $\hat{\sigma}_\alpha^2$. This is clearly unsatisfactory.

Koch (1968) shows how to overcome this difficulty. He suggests that, instead of using symmetric sums of products, one should use symmetric sums of differences. Thus in the one-way classification,

$$\begin{aligned} E(y_{ij} - y_{i'j'})^2 &= 2\sigma_e^2 && \text{when } i = i' \quad \text{and} \quad j \neq j'; \\ &= 2\left(\sigma_e^2 + \sigma_\alpha^2\right) && \text{when } i \neq i'. \end{aligned}$$

Therefore, we derive estimators from

$$2\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}\sum_{j'\neq j}^{n_i}(y_{ij}-y_{ij'})^2}{\sum_{i=1}^{a} n_i(n_i-1)} \tag{92}$$

and

$$2(\hat{\sigma}_e^2+\hat{\sigma}_\alpha^2) = \frac{\sum_{i=1}^{a}\sum_{i'\neq i}^{a}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}}(y_{ij}-y_{i'j'})^2}{\sum_{i=1}^{a}\sum_{i'\neq i}^{a} n_i n_{i'}}.$$

The resulting estimators have variances free of $\mu$ because (92) contains no terms in $\mu$. The estimators are unbiased and for balanced data, reduce to the analysis of variance estimators.

Koch (1967b) gives a procedure for obtaining an unbiased estimator of $\mu$ from an unbiased estimator of $\mu^2$. It is a by-product of (89). Suppose that our estimator of $\mu^2$ is $\hat{\mu}^2 = q(y)$, a quadratic function of $y$. This, for example, is the case for the expression in (89). Then, because $q(y)$ is unbiased for $\mu^2$, we have that

$$E(\hat{\mu}^2) = E(q(y)) = \mu^2.$$

From this, we can show for scalars $\theta$ and $g$,

$$E[q(y+\theta\mathbf{1})] = q(y) + 2g\theta + \theta^2 = \hat{\mu}^2 + 2g\theta + \theta^2.$$

Minimizing this with respect to $\theta$ gives $\theta = -g$. The minimum value is $\hat{\mu}^2 - g^2$. This suggests taking $\hat{\mu} = g$, that is, taking the estimator of $\mu$ as half the coefficient of $\theta$ in $q(y+\theta\mathbf{1})$, where $\hat{\mu}^2 = q(y)$ is derived when estimating variance components. We see that this gives the unbiased estimator

$$\hat{\mu} = g = \frac{[q(\mathbf{y}+\theta\mathbf{1}) - q(\mathbf{y}) - \theta^2]}{2\theta}.$$

Observe that

$$E(\hat{\mu}) = \frac{[(\mu+\theta)^2 - \mu^2 - \theta^2]}{2\theta} = \mu.$$

For example, we have in (89) an estimator of $\mu^2$ which is

$$q(y) = \frac{\left(y_{..}^2 - \sum_{i=1}^{a} y_{i.}^2\right)}{(N^2 - S_2)}.$$

Thus,

$$q(\mathbf{y} + \theta\mathbf{1}) = \frac{\left[(y_{..} + N\theta)^2 - \sum_{i=1}^{a} (y_{i.} + n_i\theta)^2\right]}{N^2 - S_2},$$

from which the estimator of $\mu$, taken as half the coefficient of $\theta$, is

$$\hat{\mu} = \frac{\left(Ny_{..} - \sum_{i=1}^{a} n_i y_{i.}\right)}{(N^2 - S_2)}.$$

Of course, this estimator reduces to $\bar{y}_{..}$ for balanced data.

## 7. INFINITELY MANY QUADRATICS

If the reader has gained an impression from the preceding sections that there are many quadratic forms of the observations that can be used for estimating variance components, then he has judged the situation correctly. There are infinitely many quadratic forms that can be used in the manner of the analysis of variance method. This consists of equating observed values of quadratic forms to their expected values and solving the resulting equations to get estimators of the variance components. As we have seen, this procedure is widely used but it has a serious deficiency. It gives no criteria for selecting the quadratic forms to be used. The only known property that the method gives to the resulting estimators is that they are universally unbiased for random models and, with the fitting constants method, unbiased for mixed models.

Even the property of unbiasedness is of questionable value. As a property of estimators, it has been borrowed from fixed-effects estimation. However, in the context of variance component estimation, it may not be appropriate. In estimating fixed effects, the basis of desiring unbiasedness is the concept of repetition of data and associated estimates. This basis is often not valid with unbalanced data from random models. It might perhaps be valid for repeated data. However, it might not necessarily be valid with the same pattern of unbalancedness or with the same set of (random) effects in the data. Therefore, replications of data are not just replications of any existing data structure. Mean unbiasedness may no longer be pertinent. One might consider replacing it with some other criterion. One possibility is modal unbiasedness

suggested by Searle (1968). However, Harville (1969b) doubts if modally unbiased estimators exist, and questions the justification of such a criterion on decision theoretic grounds. Nevertheless, as Kempthorne (1968) points out, mean unbiasedness in estimating fixed effects "… leads to residuals which do not have systematic effects and is therefore valuable … and is fertile mathematically in that it reduces the class of candidate statistics (or estimates)." However, "… in the variance component problem it does not lead to a fertile smaller class of estimates."

Lehmann and Casella (1998) give a general formulation of the concept of unbiasedness. Perhaps some special kind of unbiasedness other than mean or modal unbiasedness might prove useful. This needs further investigation. In recent years, there has been much work done on shrinkage estimators for means and variances. Perhaps one ought to consider these more in the context of estimating variance components. A recent source containing information on shrinkage estimators for covariance matrices is Pourahmadi (2013).

All of the estimation methods that have been discussed reduce to the analysis of variance method when the data are balanced. This and unbiasedness are the only known properties of the methods. Otherwise, the quadratic forms involved in each method have been selected solely because they seemed "reasonable" in one way or another. However, "reasonableness" of the quadratic forms in each case provides little or no comparison of any properties of the estimators that result from the different methods. Probably the simplest idea would be to compare sampling variances. Unfortunately, this comparison becomes bogged down in algebraic complexity. The variances are generally not tractable unless we assume normality. Furthermore, just as with balanced data, the variances themselves are functions of the variance components. The complexity of the variance components is evident in (63). Aside from $v(\hat{\sigma}_e^2) = 2\sigma_e^4/(N-s)$, (63) is the simplest example of a variance component estimator obtained from unbalanced data. Suppose we rewrite (63) as

$$v(\hat{\sigma}_\alpha^2) = \frac{2N}{N - \sum n_i^2} \left\{ \frac{N(N-1)(a-1)\sigma_e^4}{N^2 - \sum n_i^2} + 2\sigma_e^2\sigma_\alpha^2 + \frac{\left[N^2 \sum n_i^2 + \left(\sum n_i^2\right)^2 - 2N \sum n_i^3\right] \sigma_\alpha^4}{N\left(N^2 - \sum n_i^2\right)} \right\}. \tag{93}$$

The study of the variance in (93) as a function of $N$, the total number of observations, of $a$, the number of classes $n_i$, the number of observations in the $i$th class for $i = 1, 2, \ldots, a$, and of $\sigma_\alpha^2$ and $\sigma_e^2$ is not a small task. It would also be very difficult to compare (93) with an equally as complex a function that represents the variance of another estimator.

**TABLE 10.4 Values of $n_{ij}$ in Some 6 × 6 Designs Used by Bush and Anderson (1963)**

| Design Number | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S22 | | | | | | | C18 | | | | | | | 124 | | | | |
| 2 | 1 | 0 | 0 | 0 | 0 | | 1 | 1 | 1 | 0 | 0 | 0 | | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 2 | 1 | 0 | 0 | 0 | | 1 | 1 | 1 | 0 | 0 | 0 | | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 2 | 1 | 0 | 0 | | 0 | 1 | 1 | 1 | 0 | 0 | | 2 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 2 | 1 | 0 | | 0 | 0 | 1 | 1 | 1 | 0 | | 1 | 2 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 2 | 1 | | 0 | 0 | 0 | 1 | 1 | 1 | | 1 | 1 | 2 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 1 | 2 | | 0 | 0 | 0 | 1 | 1 | 1 | | 1 | 1 | 2 | 1 | 1 | 1 |

This is the simplest example of unbalanced data. The above discussion illustrates how analytic comparisons of the variances of different estimators presents great difficulties.

Due to the analytical difficulties just described, comparisons in the literature have largely been in terms of numerical studies. These though, are not without their difficulties also, and results can be costly to obtain. Kussmaul and Anderson (1967) have studied a special case of the two-way nested classification that makes it a particular form of the one-way classification. A study of the two-way nested classification by Anderson and Crump (1967) suggests that for very unbalanced data, the unweighted means estimator appears to have larger variance than does the analysis of variance estimator for small values of $\rho = \sigma_\alpha^2/\sigma_e^2$, but that it has smaller variance for large $\rho$. Bush and Anderson (1963) studied the two-way classification interaction model in terms of several cases of what can be called planned unbiasedness. For example, in the case of six rows and six columns, three of the designs used are those shown in Table 10.4. Designs such as these were used to compare the analysis of variance, the fitting constants and the weighted means method of estimation. Comparisons were made, by way of variances of the estimators, both of different designs as well as different estimation procedures, over a range of values of the underlying variance components. For the designs used, the general trend of the results is that, for values of the error variance much larger than the other components, the analysis of variance method estimators have smallest variance. Otherwise, the fitting constants method estimators have the smallest variance.

Even with present-day computing facilities, making comparisons such as those made by Bush and Anderson (1963) is no small task. Nevertheless, as samples of unbalanced data, generally, the examples they used (their designs) are of somewhat limited extent. This, of course, is the difficulty with numerical comparisons. Formulating planning sets of $n_{ij}$-values that will provide comparisons about unbalanced data in general is quite troublesome to say the least. Even in the simplest case, the one-way classification, there are infinitely many sets of $n_i$ values available for (93) for studying the behavior of $v(\hat{\sigma}_\alpha^2)$. In addition, there are varying values of $a$ and of $\sigma_\alpha^2$ and $\sigma_e^2$ that one has to take into consideration. There is considerable difficulty in planning a series of these values that in any sense "covers the field." This difficulty is simply multiplied when one comes to consider higher order classifications, such as those handled by Bush and Anderson (1963). Therefore, neither analytic nor

numerical comparisons of estimators are easily resolved. One thing that can be done is to go back to the grounds on which "reasonableness" was judged appropriate in establishing the methods. Searle (1971b) summarizes the situation. "The analysis of variance method commends itself because it is the obvious analogue of the analysis of variance of balanced data and is easy to use; some of its terms are not sums of squares, and it gives biased estimators in mixed models." The generalized form of Henderson's method 2 makes up for this deficiency, but it is not uniquely defined and his specific definition cannot be used when there are interactions between fixed and random effects. The fitting constants method uses sums of squares that have non-central $\chi^2$-distributions in the fixed-effects models, and it gives unbiased estimators in the mixed model; but it can involve more quadratics that there are components to be estimated; and it can also involve extensive computing" (inverting matrices of order equal to the number of random effects in the model). For data in which all subclasses are filled the analysis of means, methods have the advantage of being easier to compute than the fitting constants method. The unweighted means analysis is especially easy. For balanced data, all of them reduce to the analysis of variance method. Moreover, all of them can yield negative estimates for the variance components. Little more can be said by way of comparing the methods. The problem awaits thorough investigation.

## 8. MAXIMUM LIKELIHOOD FOR MIXED MODELS

We have already mentioned (see Section 2 of Chapter 9) that all models could, in fact, be called mixed models. Every model usually has both a general mean $\mu$, which is a fixed effect and error terms $\mathbf{e}$ which are random. Thus, although by its title this section might be devoted to only one class of models it does in fact apply to all linear models.

The fitting constants method of estimating variance components gives unbiased estimators of the components even for mixed models. However, it is only a method for estimating variance components of the model and gives no guidance on the problem of estimating the fixed effects. If the variance components of the model are known, of course, there would be no problem in estimating the estimable functions of the fixed effects from a solution to the normal equations $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{b}^{\circ} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ of the generalized least-square procedure. In these equations, $\mathbf{V}$ is the variance–covariance matrix of $\mathbf{y}$. The elements of $\mathbf{V}$ are functions of the (assumed known) variance components. However, when these variance components are unknown, as is usually the case, we want to be able to estimate both the fixed effects and the variance components of the model simultaneously.

At least two courses of action are available. They include the following:

(i) Use the fitting constants method to estimate the variance components. Then, use the resulting estimates in place of the true components in $\mathbf{V}$ in the generalized least-square equations for the fixed effects.

(ii) Estimate the fixed effects and the variance components simultaneously with a unified procedure such as maximum likelihood.

In both cases, recourse usually has to be made to an iterative procedure with its possibly extensive computing requirements. However, some progress has been made analytically. We shall indicate the results of some of this progress.

### a. Estimating Fixed Effects

We write the model,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}. \tag{94}$$

The fixed effects are contained in the vector **b.** The vector **u** contains the random effects. The corresponding incidence matrices are **X** and **Z**. The vector **e** contains the random error terms. In the usual way, we assume that the error terms have zero means, are uncorrelated and in this case have known variance–covariance matrices

$$\text{var}(\mathbf{u}) = E(\mathbf{uu}') = \mathbf{D} \quad \text{and} \quad \text{var}(\mathbf{e}) = E(\mathbf{ee}') = \mathbf{R}. \tag{95}$$

From (94), it follows that

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R}. \tag{96}$$

We also assume that **V** is non-singular. The normal equations stemming from generalized least squares are

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{Xb}^{\circ} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{97}$$

with solution

$$\mathbf{b}^{\circ} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{98}$$

If **V** is singular, we replace $\mathbf{V}^{-1}$ in (97) and (98) by $\mathbf{V}^{-}$ as in (143a) and (143b) of Chapter 5. Under normality assumptions for the $u$'s and the $e$'s, (98) also represents the maximum likelihood solution.

Calculating (98) involves $\mathbf{V}^{-1}$. The order of this matrix is equal to the number of observations. For large data sets this can be very large, perhaps many thousands. After we have obtained $\mathbf{V}^{-1}$, we also need a generalized inverse $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$. Obtaining this generalized inverse will be a lesser task because its order is the number of levels of the fixed effects. Therefore, the difficulty with (97) or (98) is that of calculating $\mathbf{V}^{-1}$. In the fixed effects case, **V** usually has the form $\sigma_e^2\mathbf{I}_N$. With a little more generality, it may be diagonal. In either case, inversion of **V** is simple. However, in general $\mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}$ of (96) is not diagonal even if **D** and **R** are. Thus it is not always easy to calculate $\mathbf{V}^{-1}$. However, Henderson et al. (1959) give an alternative derivation of $\mathbf{b}^{\circ}$ by establishing a set of equations that do not involve $\mathbf{V}^{-1}$. We now show how he does this.

Suppose that in (94), the effects represented by a vector $\mathbf{u}$ were in fact fixed and not random. Then, because $\text{var}(\mathbf{e}) = \mathbf{R}$, the normal equations for the now completely fixed-effects model would be

$$\begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} \mathbf{R}^{-1} \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} \mathbf{R}^{-1}\mathbf{y}$$

or, equivalently,

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{99}$$

We use the notation $\tilde{\mathbf{b}}$ in contrast to $\mathbf{b}^\circ$ to distinguish a solution to (99) from one of (97).

Suppose that we amend (99) by adding $\mathbf{D}^{-1}$ to the lower right-hand sub-matrix $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$ of the matrix on the left. This gives

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{100}$$

We use the asterisk notation to distinguish solutions of (100). We can show that the solutions $\mathbf{b}^*$ to (100) are identical to the solutions $\mathbf{b}^\circ$ of (97). In this way, (100) provides a means of deriving $\mathbf{b}^\circ$ without having to invert $\mathbf{V}$. We only have to invert $\mathbf{D}$ and $\mathbf{R}$. These matrices are usually diagonal. We then have to solve (100) which has as many equations as there are fixed and random effects in the model. Usually, this is considerably fewer than the number of observations, so (100) is easier to solve than (97).

We now demonstrate the equivalence of $\mathbf{b}^*$ of (100) to $\mathbf{b}^\circ$ of (98). From (100), observing that for the second rows of the matrices

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X})\mathbf{b}^* + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{u}^* = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}$$

and solving for $\mathbf{u}^*$, we get

$$\mathbf{u}^* = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\mathbf{b}^*).$$

Substituting $\mathbf{u}^*$ into the system of equations in the first row of the matrices, we get

$$(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})\mathbf{b}^* + (\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z})(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\mathbf{b}^*) = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}.$$

Thus, putting terms involving $\mathbf{b}^*$ on the left-hand side of the equation and terms involving $\mathbf{y}$ on the right-hand side, we have,

$$\begin{aligned} &\mathbf{X}'[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}]\mathbf{X}\mathbf{b}^* \\ &\quad = \mathbf{X}'[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}]\mathbf{y}. \end{aligned} \tag{101a}$$

Substituting

$$\mathbf{W} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$$

into (101a) yields

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^* = \mathbf{X}'\mathbf{W}\mathbf{y}. \tag{101b}$$

However,

$$\begin{aligned}\mathbf{W}\mathbf{V} &= [\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}](\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}) \\ &= \mathbf{R}^{-1}\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{Z}') \\ &= \mathbf{R}^{-1}\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{D}\mathbf{Z}' \\ &= \mathbf{I}.\end{aligned}$$

Thus, $\mathbf{W} = \mathbf{V}^{-1}$. Therefore, equations (101) and (97) are the same. As a result the solution $\mathbf{b}^*$ to (101b), which in part is a solution to (100), is the solution to (97) given in (98). Therefore, equation (100), with its computational advantages over (97) can be used to derive a solution to (97).

Equations (100) are the normal equations of the model assuming that all effects are fixed. They are equations (99) modified by adding the inverse of the variance–covariance matrix of the random effects $\mathbf{u}$ to the sub-matrix that is the coefficient of $\tilde{\mathbf{u}}$ in the "$\tilde{\mathbf{u}}$-equations"–, that is, by adding $\mathbf{D}^{-1}$ to $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$, as in (100). In certain special cases, this is particularly simple. For example, when $\mathbf{R} = \text{var}(\mathbf{e}) = \sigma_e^2\mathbf{I}_N$, as is so often, assumed equations (99) are

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{b}} \\ \tilde{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y}\end{bmatrix} \tag{102}$$

and equations (100) are

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2\mathbf{D}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{b}^* \\ \mathbf{u}^*\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y}\end{bmatrix}. \tag{103}$$

Furthermore, $\mathbf{D}$ is often diagonal of the form

$$\mathbf{D} = \text{diag}\{\sigma_\theta^2\mathbf{I}_{N_\theta}\} \quad \text{for} \quad \theta = A, B, \ldots, K,$$

where $A, B, \ldots, K$ are the random factors, the factor $\theta$ having $N_\theta$ levels and variance $\sigma_\theta^2$.

In this case, $\sigma_\theta^2\mathbf{D}^{-1}$ requires just adding $\sigma_e^2/\sigma_\theta^2$ to appropriate diagonal elements of $\mathbf{Z}'\mathbf{Z}$. In particular, if there is only one random factor, (103) becomes

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \left(\dfrac{\sigma_e^2}{\sigma_\theta^2}\right)\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}. \tag{104}$$

Of course, this formulation of the maximum likelihood solution $\mathbf{b}^\circ = \mathbf{b}^*$ applies only when the variance components are known. However, in most applications just their values relative to $\sigma_e^2$ need be known. The result in (104) illustrates this. However, together with the fitting constants method of estimating variance components free of the fixed effects, (100) and its simplified forms provide a framework for estimating both the fixed effects and the variance components of a mixed model.

Equations (100) arise from the joint density of $\mathbf{y}$ and $\mathbf{u}$. Assuming that $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$, this joint density is

$$\begin{aligned} f(\mathbf{y}, \mathbf{u}) &= g(\mathbf{y}|\mathbf{u})h(\mathbf{u}) \\ &= C \exp\left[-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{u})\right] \exp\left[-\tfrac{1}{2}\mathbf{u}'\mathbf{D}\mathbf{u}\right], \end{aligned}$$

where $C$ is a constant. Maximizing with respect to $\mathbf{b}$ and $\mathbf{u}$ leads at once to (100).

The solution for $\mathbf{b}^*$ in (100) is of interest because $\mathbf{b}$ is a vector of fixed effects in the model (94). However, even though $\mathbf{u}$ is a vector of random variables, the solution for $\mathbf{u}^*$ in (100) is, in many situations, also of interest. It is an estimator of the conditional mean of $\mathbf{u}$ given $\mathbf{y}$. We now show this. First, from (94) and (95), we have $\text{cov}(\mathbf{u}, \mathbf{y}') = \mathbf{D}\mathbf{Z}'$. Then, on assuming normality,

$$E(\mathbf{u}|\mathbf{y}) = E(\mathbf{u}) + \text{cov}(\mathbf{u}, \mathbf{y}')[\text{var}(\mathbf{y})]^{-1}[\mathbf{y} - E(\mathbf{y})] = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

Hence, from (100),

$$\begin{aligned} \mathbf{u}^* &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{V}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^*). \end{aligned} \tag{105}$$

This last expression in the above equation is exactly $E(\mathbf{u}|\mathbf{y})$ with $\mathbf{b}$ replaced by $\mathbf{b}^*$, which we know is the maximum likelihood estimator of $\mathbf{b}$. Hence, for a given set of observations $\mathbf{y}$, $\mathbf{u}^* = \widehat{E(\mathbf{u}|\mathbf{y})}$ is the maximum likelihood estimator of the mean of $\mathbf{u}$.

Henderson et al. (1959) mentions with further discussion in Henderson (1963) that $\mathbf{u}^* = \widehat{E(\mathbf{u}|\mathbf{y})}$ is the "estimated genetic merit" as used by animal breeders. In their

case, $\mathbf{u}$ is a vector of genetic merit values of a series of animals from which $\mathbf{y}$ is the vector of production records. The problem is to use $\mathbf{y}$ to get estimated values of $\mathbf{u}$ in order to determine which animals are best in some sense. For example, if the animals were cows, $\mathbf{y}$ might represent how many gallons of milk they produce in a week.

The estimators $\mathbf{b}^*$ and $\mathbf{u}^*$ derived above are often referred to in the literature as the best linear unbiased predictor (BLUP). The original presentation of this estimator is due to Henderson. Another derivation of this estimator using the linear Bayes estimator (see Section 3 of Chapter 3), also available in Gruber (2010) modeled after that of Bulmer (1980), will now be given. Two other useful references are Henderson (1975) and Robinson (1991).

Consider the linear model where the vector $\mathbf{b}$ contains the fixed effects and $\mathbf{u}$ is a random vector.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{106}$$

Assume that

$$E(\mathbf{u}) = \mathbf{0} \text{ and } D(\mathbf{u}) = \mathbf{D}.$$

Also, assume that

$$E(\mathbf{e}) = \mathbf{0} \text{ and } D(e) = \mathbf{R}.$$

Let $\mathbf{v} = \mathbf{Y} - \mathbf{Xb}$. Then, for the model

$$\mathbf{v} = \mathbf{Zu} + \mathbf{e},$$

the linear Bayes estimator of $\mathbf{u}$ takes the form

$$\tilde{\mathbf{u}} = \mathbf{DZ}'(\mathbf{ZDZ}' + \mathbf{R})^{-1}\mathbf{v} \tag{107}$$

Rewrite the model in (106) as

$$\mathbf{y} = \mathbf{Xb} + \eta, \tag{108}$$

where $\eta = \mathbf{Zu} + \mathbf{e}$. It follows that

$E(\eta) = \mathbf{0}$ and $D(\eta) = \mathbf{XDX}' + \mathbf{R} = \mathbf{W}$. Then, the weighted least-square estimator for $\mathbf{b}$ is

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \tag{109}$$

Substitution of (109) into (107) yields

$$\mathbf{u}^* = \mathbf{DZ}'(\mathbf{ZDZ}' + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{Xb}^*). \tag{110}$$

This is equation (105) with $\mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}$. Thus, (110) is the BLUP.

### b. Fixed Effects and Variance Components

We cannot solve maximum likelihood equations for estimating variance components for unbalanced data explicitly. The equations for the simplest case illustrate this. Consider the one-way classification as described in Section 6. With

$$\mathbf{V} = \text{var}(\mathbf{y}) = \sigma_e^2 \mathbf{I}_N + \sigma_\sigma^2 \sum_{i=1}^{a} {}^{+}\mathbf{J}_{n_i},$$

as used there,

$$|\mathbf{V}| = \sigma_e^{2(N-a)} \prod_{i=1}^{a} \left(\sigma_e^2 + n_i \sigma_\alpha^2\right),$$

and

$$\mathbf{V}^{-1} = \left(\frac{1}{\sigma_e^2}\right)\mathbf{I}_N + \sum_{i=1}^{a} {}^{+}\frac{1}{n_i}\left(\frac{1}{\sigma_e^2 + n_i\sigma_\alpha^2}\right)\mathbf{J}_{n_i}.$$

On the basis of normality, the likelihood function is

$$L = (2\pi)^{-\frac{1}{2}N} |\mathbf{V}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{y} - \mu\mathbf{1})'(\mathbf{y} - \mu\mathbf{1})\right\},$$

and after substituting for $|\mathbf{V}|$ and $\mathbf{V}^{-1}$, the logarithm of this reduces to

$$\log L = \tfrac{1}{2}N\log(2\pi) - \tfrac{1}{2}(N-a)\log\sigma_e^2 - \tfrac{1}{2}\sum_{i=1}^{a}\log\left(\sigma_e^2 + n_i\sigma_\alpha^2\right)$$
$$- \tfrac{1}{2}\left(\frac{1}{\sigma_e^2}\right)\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 - \tfrac{1}{2}\sum_{i=1}^{a}\frac{n_i(\bar{y}_{i.} - \mu)^2}{\sigma_e^2 + n_i\sigma_\alpha^2}.$$

Equating to zero, the differentials of $\log L$ with respect to $\mu$, $\sigma_e^2$ and $\sigma_\alpha^2$, gives, formally, the equations whose solutions (to be denoted by $\tilde{\mu}, \tilde{\sigma}_e^2$ and $\tilde{\sigma}_\alpha^2$) are the maximum likelihood estimators. These equations are as follows:

$$\tilde{\mu} = \frac{\displaystyle\sum_{i=1}^{a}\frac{n_i\bar{y}_{i.}}{\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2}}{\displaystyle\sum_{i=1}^{a}\frac{n_i}{\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2}},$$

$$\frac{N-a}{\tilde{\sigma}_e^2} + \sum_{i=1}^{a}\frac{1}{\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2} - \frac{\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{\tilde{\sigma}_e^4} - \sum_{i=1}^{a}\frac{n_i(\bar{y}_{i.} - \tilde{\mu})^2}{\left(\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2\right)^2} = 0,$$

and

$$\sum_{i=1}^{a} \frac{n_i}{\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2} - \sum_{i=1}^{a} \frac{n_i^2(\bar{y}_{i.}^2 - \tilde{\mu})^2}{\left(\tilde{\sigma}_e^2 + n_i\tilde{\sigma}_\alpha^2\right)^2} = 0.$$

These equations have no explicit solution for $\tilde{\mu}, \tilde{\sigma}_e^2$, and $\tilde{\sigma}_\alpha^2$. Of course, they reduce to the simpler equations of balanced data given in (77) of Section 9g of Chapter 9, when $n_i = n$ for all $i$.

Even if we could find solutions in the unbalanced data case, we still must consider the problem of using these solutions to derive a non-negative estimator of $\sigma_\alpha^2$. We had the same consideration for the balanced case at the end of Section 9g of Chapter 9.

Therefore, explicit maximum likelihood estimators must be despaired of. However, Hartley and J. N. K. Rao (1967) have developed a general set of equations from which specific estimates are obtained by iteration involving extensive computations. We give their equations and mention how they indicate a solution may be found. To do so, we rewrite the model (94) using

$$\mathbf{Zu} = \sum_{\theta=A}^{K} \mathbf{Z}_\theta \mathbf{u}_\theta,$$

where $\mathbf{u}_\theta$ is the vector of random effects of the $\theta$-factor. Then, defining $\gamma_\theta$ as $\gamma_\theta = \sigma_\theta^2/\sigma_e^2$ for $\theta = A, B, \ldots, K$, and $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{I}_N + \sum_{\theta=A}^{K} \gamma_\theta \mathbf{Z}_\theta \mathbf{Z}_\theta', \tag{111}$$

$\mathbf{V}$ of (96) is $\mathbf{V} = \sigma_e^2\mathbf{H}$. On assuming normality, the logarithm of the likelihood is

$$\log L = \tfrac{1}{2}N\log(2\pi) - \tfrac{1}{2}N\log\sigma_e^2 - \frac{(\mathbf{y} - \mathbf{Xb})'\tilde{\mathbf{H}}^{-1}(\mathbf{y} - \mathbf{Xb})}{2\sigma_e^2}.$$

Equating the differentials of $\log L$ with respect to $\sigma_e^2$, the $\gamma_\theta$ and the elements of $\mathbf{b}$ to zero gives the following equations:

$$\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{Xb} = \mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{y}, \tag{112}$$

$$\sigma_e^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})'\tilde{\mathbf{H}}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})}{N}, \tag{113}$$

and

$$\mathrm{tr}(\tilde{\mathbf{H}}^{-1}\mathbf{Z}_\theta\mathbf{Z}_\theta') = \frac{(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})'\tilde{\mathbf{H}}^{-1}\mathbf{Z}_\theta\mathbf{Z}_\theta'\tilde{\mathbf{H}}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}})}{\tilde{\sigma}_e^2} \text{ for } \theta = A, B, \ldots, K. \tag{114}$$

These equations have to be solved for the elements of $\tilde{\mathbf{b}}$, the error variance $\sigma_e^2$, and the variance components inherent in $\tilde{\mathbf{H}}$. Hartley and Rao (1967) indicate how this can be done, either by the method of steepest ascent or obtaining an alternative form for (114). However, the equations for the alternative form of (114) are difficult to handle. Of course, equations (112) and (113) are recognizable as the maximum likelihood equations for the fixed effects and the error variance. They are easy to solve if values of the $\tilde{\gamma}_\theta$'s are available for $\tilde{\mathbf{H}}$. Thus, an iteration is established via equations (112), (113), and (114).

### c. Large Sample Variances

Searle (1970) obtained general expressions for large sample variances of maximum likelihood estimators of variance components under normality assumptions. We can derive these expressions despite the fact that the estimators themselves cannot be obtained explicitly. Using the model (94), with $\text{var}(\mathbf{Zu} + \mathbf{e}) = \mathbf{V}$ as in (96) and with $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$, the likelihood of the sample is

$$L = (2\pi)^{-\frac{1}{2}N}|\mathbf{V}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})\right\}.$$

The logarithm of this likelihood is

$$\log L = -\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}). \tag{115}$$

Suppose the model has $p$ fixed effects and $q$ variance components represented by $\boldsymbol{\sigma}^2 = \{\sigma_i^2\}$ for $i = 1, 2, \ldots, q$, one element of $\boldsymbol{\sigma}^2$ being $\sigma_e^2$. Then (see Wald (1943)), the variance–covariance matrix for the large sample maximum likelihood estimators of the $p$ elements of $\mathbf{b}$ and the $q$ variance components is

$$\begin{bmatrix} \text{var}(\tilde{\mathbf{b}}) & \text{cov}(\tilde{\mathbf{b}}, \tilde{\sigma}^2) \\ \text{cov}(\tilde{\sigma}^2, \tilde{\mathbf{b}}) & \text{var}(\tilde{\sigma}^2) \end{bmatrix} = \begin{bmatrix} -E(\mathbf{L}_{bb}) & -E(\mathbf{L}_{b\sigma^2}) \\ -E(\mathbf{L}_{b\sigma^2})' & -E(\mathbf{L}_{\sigma^2\sigma^2}) \end{bmatrix}^{-1}. \tag{116}$$

In (116), $\tilde{\mathbf{b}}$ and $\tilde{\sigma}^2$ are the maximum likelihood estimators of $\mathbf{b}$ and $\sigma^2$, respectively. The left-hand side of (116) is a statement of their covariance matrix. The right-hand side shows how to derive this covariance matrix. In its sub-matrices $\mathbf{L}_{bb}$, for example, is the $p \times p$ matrix of second differentials of $\mathbf{L}$ of (115) with respect to the elements of $\mathbf{b}$. The definition of $L_{b\sigma^2}$ and $L_{\sigma^2\sigma^2}$ follows in a similar manner.

The nature of (115) is such that after some algebraic manipulations, (116) yields the following results:

$$\text{var}(\tilde{\mathbf{b}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{117}$$

$$\text{cov}(\tilde{\mathbf{b}}, \tilde{\sigma}^2) = \mathbf{0}, \tag{118}$$

and

$$\text{var}(\tilde{\sigma}^2) = 2\left\{\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_j^2}\right) \text{ for } i,j = 1\cdots q\right\}^{-1}. \tag{119}$$

Searle (1970) gives details of the derivation of these results.

The results (117)–(119) merit attention. First, (117) corresponds to the variance of $\mathbf{b}^\circ$ in (98) and therefore comes as no surprise. Nevertheless, it indicates that for unbalanced data from any mixed model, the maximum likelihood (under normality) estimators of the fixed effects is what it would be if the variance components were known and did not have to be estimated. Second, (118) shows that the covariance between large sample maximum likelihood estimators of fixed effects and variance components are zero. The simplest case of this relates to the mean of a sample and the sample variance. Under normality, they are distributed independently. The generalization of this result is (118), which is therefore no surprise either. Finally, (119) gives the variance–covariance matrix of the large sample maximum likelihood estimators of the variance components. We notice that it is quite free of $\mathbf{X}$, the incidence matrix of the fixed effects. As one can see from (119) its form is the inverse of a matrix whose typical element is the trace of the product of matrices $\mathbf{V}^{-1}$ and the derivatives of $\mathbf{V}$ with respect to the variance components.

**Example 7 Variance of a Variance Component in a Simple Case** Consider $N$ observations from the model $y_i = \mu + e_i$ with $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Then, $\mathbf{V} = \sigma^2\mathbf{I}_N$, $\mathbf{V}^{-1} = (1/\sigma^2)\mathbf{I}_N$ and $\mathbf{V}_{\sigma^2} = \mathbf{I}_N$. Hence, from (119), we obtain the well-known result

$$\text{var}(\tilde{\sigma}^2) = 2\left\{\left[\text{tr}\left(\tfrac{1}{\sigma^2}\right)\mathbf{I}_N\mathbf{I}_N\right]^2\right\}^{-1} = 2\left(\tfrac{N}{\sigma^4}\right)^{-1} = \tfrac{2\sigma^4}{N}.$$

□

Additional results stemming from (119) are shown in the next chapter on the web page.

## 9. MIXED MODELS HAVING ONE RANDOM FACTOR

The mixed model (94) has several simplifying features when it has only one factor that is random. We assume that in

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \tag{120}$$

$r(\mathbf{X}) = r$, with $\mathbf{b}$ representing $q \geq r$ fixed effects and $\mathbf{u}$, in representing the random effects, contains $t$ effects for just one random factor having variance $\sigma_u^2$. As a result, $\mathbf{Z}$ has full-column rank, $t$, with its columns summing to 1, the same, as do certain columns of $\mathbf{X}$. We assume that this is the only linear relationship of the columns of

$\mathbf{Z}$ to those of $\mathbf{X}$. Hence,

$$r\begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix} = r(\mathbf{X}) + t - 1 = r + t - 1.$$

Furthermore, by the nature of $\mathbf{Z}$, the matrix $\mathbf{Z}'\mathbf{Z}$ is diagonal of order $t$ and $(\mathbf{Z}'\mathbf{Z})^{-1}$ exists.

Since the model is a mixed model, estimation is by the fitting constants method, using

$$\text{SSE} = \mathbf{y}'\mathbf{y} - R(\mathbf{b}, \mathbf{u}) \quad \text{and} \quad R(\mathbf{u}|\mathbf{b}) = R(\mathbf{b}, \mathbf{u}) - R(\mathbf{b})$$

with

$$E(\text{SSE}) = [N - (r + t - 1)]\sigma_e^2 \tag{121}$$

in the usual manner. From (79),

$$E[R(\mathbf{u}|\mathbf{b})] = \sigma_u^2 \text{tr}[\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}] + \sigma_e^2[r(\mathbf{X}) + t - 1 - r(\mathbf{X})]. \tag{122}$$

Hence, the estimators are

$$\hat{\sigma}_e^2 = \frac{\mathbf{y}'\mathbf{y} - R(\mathbf{b}, \mathbf{u})}{N - r(\mathbf{X}) - t + 1} \tag{123}$$

and

$$\hat{\sigma}_u^2 = \frac{R(\mathbf{u}|\mathbf{b}) - \hat{\sigma}_e^2(t - 1)}{\text{tr}[\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}]}. \tag{124}$$

These are the estimators given by Cunningham and Henderson (1968) for the case where $\mathbf{X}$ has full-column rank. For a particular case of ensuring the non-singularity of $\mathbf{X}'\mathbf{X}$ through appropriate constraints, Cunningham (1969) gives a simple expression for the denominator of (124).

A problem with the preceding formulation is that the calculation of

$$R(\mathbf{b}, \mathbf{u}) = \mathbf{y}' \begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix} \begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\end{bmatrix} \begin{bmatrix}\mathbf{X} \\ \mathbf{Z}\end{bmatrix} \mathbf{y} \tag{125}$$

may be very difficult because $\mathbf{Z}$ has as many columns as there are random effects in the data. Since the random effects can be quite numerous, the computation of (125) could be onerous. However, a generalization of the "absorption process" described in Chapter 7 for the two-way classification permits an easier calculation as follows. With the easier-to-calculate quantity

$$R(\mathbf{u}) = \mathbf{y}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \tag{126}$$

we find that

$$R(\mathbf{b}|\mathbf{u}) = R(\mathbf{b}, \mathbf{u}) - R(\mathbf{u}) \tag{127}$$

simplifies, after substitution from (125) and (126) to

$$R(\mathbf{b}|\mathbf{u}) = \mathbf{b}^{\circ\prime}\mathbf{X}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y}, \tag{128}$$

where

$$\mathbf{b}^{\circ} = \mathbf{Q}^{-}\mathbf{X}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y} \tag{129}$$

with

$$\mathbf{Q} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}. \tag{130}$$

The quantity in (128) is easier to compute than that of (125) because $\mathbf{Q}$ and $\mathbf{b}^{\circ}$ have $q$ rows. Using (126) and (128), we then calculate $R(\mathbf{b}, \mathbf{u})$ as

$$R(\mathbf{b}, \mathbf{u}) = R(\mathbf{b}|\mathbf{u}) + R(\mathbf{u}).$$

Hence, for (124), we calculate $R(\mathbf{u}|\mathbf{b})$ as

$$R(\mathbf{u}|\mathbf{b}) = R(\mathbf{b}|\mathbf{u}) + R(\mathbf{u}) - R(\mathbf{b}), \tag{131}$$

where

$$R(\mathbf{b}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y} \tag{132}$$

is also easy to compute.

Results (128)–(132) are similar to those summarized by Cunningham and Henderson (1968) for a model in which $\mathbf{X}$ is assumed to have full-column rank. As we shall see, this restriction is not necessary. The crucial result is (128), derived from (127) by substituting from (125) and (126) using

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^{-} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'\mathbf{Z})^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ -(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \end{bmatrix} \mathbf{Q}^{-}[\mathbf{I} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}] \tag{133}$$

with $\mathbf{Q}$ of (130). (See Exercise 33 of Chapter 1.) When carrying out the derivation, we find that $\mathbf{b}^{\circ}$ of (129) is a solution to

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{b}^{\circ} \\ \mathbf{u}^{\circ} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}. \tag{134}$$

These are the least-square normal equations for $\mathbf{b}^\circ$ and $\mathbf{u}^\circ$ assuming that $\mathbf{u}$ is a vector fixed rather than random effects. Recall, however, that comparable equations for getting maximum likelihood solutions for the fixed effects are from (104),

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b}^* \\ \mathbf{u}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{135}$$

where

$$\lambda = \sigma_e^2/\sigma_u^2. \tag{136}$$

Since (135) is formally the same as (134) except for $\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I}$ replacing $\mathbf{Z}'\mathbf{Z}$, Cunningham and Henderson (1968) suggested making this replacement throughout the whole variance component estimation process described in (123) through (131). The result is an iterative procedure based on the maximum likelihood equations implicit in (135). Thus, (123) and (124) would become

$$\sigma_e^{*2} = \frac{\mathbf{y}'\mathbf{y} - R^*(\mathbf{b}, \mathbf{u})}{N - r(\mathbf{X}) - t + 1} \tag{137}$$

and

$$\sigma_u^{*2} = \frac{R^*(\mathbf{u}|\mathbf{b}) - \hat{\sigma}_e^{*2}(t-1)}{\mathrm{tr}[\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}]}. \tag{138}$$

The comparable definitions of the $R^*$-terms are

$$R^*(\mathbf{b}, \mathbf{u}) = R^*(\mathbf{b}|\mathbf{u}) + R^*(\mathbf{u}) \tag{139}$$

derived from using

$$\mathbf{P} = \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I} \tag{140}$$

in place of $\mathbf{Z}'\mathbf{Z}$ in (126)–(132). Thus, from (126),

$$\mathbf{R}^*(\mathbf{u}) = \mathbf{y}'\mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}'\mathbf{y}. \tag{141}$$

From (128)–(130) we have,

$$R^*(\mathbf{b}|\mathbf{u}) = \mathbf{y}'(\mathbf{I} - \mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}')'\mathbf{X}[\mathbf{X}'(\mathbf{I} - \mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}')\mathbf{X}]^-\mathbf{X}'(\mathbf{I} - \mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}')\mathbf{y}. \tag{142}$$

Equation (132) remains the same, namely,

$$R(\mathbf{b}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y}. \tag{143}$$

Then, for (138), just as in (131),

$$R^*(\mathbf{u}|\mathbf{b}) = R^*(\mathbf{b}|\mathbf{u}) + R^*(\mathbf{u}) - R^*(\mathbf{b}). \tag{144}$$

The replacement of $\mathbf{Z}'\mathbf{Z}$ by $\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I}$ as just described is based on the premise that the expected values of $\text{SSE}^* = \mathbf{y}'\mathbf{y} - R^*(\mathbf{b}|\mathbf{u})$ and $R^*(\mathbf{u}|\mathbf{b})$ are those of SSE and $R(\mathbf{u}|\mathbf{b})$ as shown in (121) and (122) with $\mathbf{Z}'\mathbf{Z}$ replaced by $\mathbf{P}$. Thompson (1969) points out that this is not the case. Consequently, (137) and (138) are not unbiased estimators. The derivation of unbiased estimators that Thompson (1969) indicates proceeds as follows. First, notice that from (140),

$$\begin{aligned}
\mathbf{P}^{-1}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2\mathbf{I}) &= \mathbf{P}^{-1}\big(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2\mathbf{I}\big)\mathbf{Z}' \\
&= \mathbf{P}^{-1}(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I})\sigma_u^2\mathbf{Z}', \text{ from (136)} \\
&= \mathbf{P}^{-1}\mathbf{P}\mathbf{Z}'\sigma_u^2, \text{ from (140)} \\
&= \mathbf{Z}'\sigma_u^2.
\end{aligned} \tag{145}$$

Second, it follows from (120) that

$$E(\mathbf{y}\mathbf{y}') = \mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + \mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2\mathbf{I}. \tag{146}$$

Hence, using $E(\mathbf{y}'\mathbf{A}\mathbf{y}) = \text{tr}[\mathbf{A}E(\mathbf{y}\mathbf{y}')]$, the expected value of (141) is

$$\begin{aligned}
E[R^*(\mathbf{u})] &= \text{tr}[\mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}'E(\mathbf{y}\mathbf{y}')] \\
&= \text{tr}\big[\mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}'\mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + \mathbf{Z}\mathbf{Z}'\sigma_u^2\big].
\end{aligned} \tag{147}$$

Similarly, if

$$\mathbf{T} = \mathbf{I} - \mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}', \tag{148}$$

(145) gives $\mathbf{T}(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2\mathbf{I}) = \sigma_e^2\mathbf{I}$. Thus, from (142) and (146), we have

$$\begin{aligned}
E[R^*(\mathbf{b}|\mathbf{u})] &= \text{tr}[\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-\mathbf{X}'\mathbf{T}E(\mathbf{y}\mathbf{y}')] \\
&= \text{tr}\big[\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-\mathbf{X}'\mathbf{T}\mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + \mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-\mathbf{X}'\sigma_e^2\big] \\
&= \text{tr}\big[\mathbf{T}\mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + \mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-\mathbf{X}'\sigma_e^2\big].
\end{aligned} \tag{149}$$

From (143) and (146), we have that

$$\begin{aligned}
E(R^*(\mathbf{b})) &= \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'E(\mathbf{y}\mathbf{y}')] \\
&= \text{tr}\big[\mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\big(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2\mathbf{I}\big)\big].
\end{aligned} \tag{150}$$

Therefore, from (144), using (147), (149), and (150),

$$\begin{aligned} E[R^*(\mathbf{u}|\mathbf{b})] &= \mathrm{tr}\big\{(\mathbf{Z}\mathbf{P}^{-1}\mathbf{Z}' + \mathbf{T} - \mathbf{I})\mathbf{X}\mathbf{b}\mathbf{b}'\mathbf{X}' + [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{Z}\mathbf{Z}'\sigma_u^2 \\ &\quad + [\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{Z}\mathbf{Z}'\sigma_e^2\big\} \\ &= \sigma_u^2\mathrm{tr}[\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}] \\ &\quad + \sigma_e^2\mathrm{tr}[\mathbf{X}'\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^- - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-]. \end{aligned} \tag{151}$$

By Lemma 1 in Section 2c of Chapter 1 $\mathbf{X}'\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-$ and $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-$ are idempotent matrices. The rank of an idempotent matrix is equal to its trace. Thus, we have that $\mathrm{tr}(\mathbf{X}'\mathbf{T}\mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^-) = r(\mathbf{X}'\mathbf{T}\mathbf{X})$ and $\mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-) = r(\mathbf{X}'\mathbf{X})$. Furthermore, $\mathbf{T}$ has full rank (its inverse being $(\mathbf{Z}'\mathbf{Z}/\lambda + \mathbf{I})$). Thus, $r(\mathbf{X}\mathbf{T}\mathbf{X}') = r(\mathbf{X})$. Hence, the last term of (151) is zero. Thus

$$E[R^*(\mathbf{u}|\mathbf{b})] = \sigma_u^2\mathrm{tr}[\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}].$$

Moreover, from (146)–(149),

$$E[\mathbf{y}'\mathbf{y} - R^*(\mathbf{b}, \mathbf{u})] = E[\mathbf{y}'\mathbf{y} - R^*(\mathbf{u}) - R^*(\mathbf{b}|\mathbf{u})] = [N - r(\mathbf{X})]\sigma_e^2.$$

Therefore, in place of (137) and (138), estimators for $\sigma_e^2$ and $\sigma_u^2$ are

$$\tilde{\sigma}_e^2 = \frac{\mathbf{y}'\mathbf{y} - [R^*(\mathbf{u}) + R^*(\mathbf{b}|\mathbf{u})]}{N - r(X)} \tag{152}$$

and

$$\tilde{\sigma}_u^2 = \frac{R^*(\mathbf{u}) + R^*(\mathbf{b}|\mathbf{u}) - R^*(\mathbf{b})}{\mathrm{tr}[\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}]}. \tag{153}$$

These results, given by Thompson (1969) for $\mathbf{X}$ of full-column rank provide an iterative procedure. This is because through $\mathbf{P}$ of (140), the reductions $R^*(\mathbf{u})$ and $R^*(\mathbf{b}|\mathbf{u})$ of (141) and (142) involve $\lambda = \sigma_e^2/\sigma_u^2$. Therefore, we achieve estimation by taking an initial value of $\lambda$, calculating (152) and (153), using the results to get a next value of $\lambda$ and repeating the process until convergence is attained.

The replacement of $\mathbf{Z}'\mathbf{Z}$ by $\mathbf{P} = \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I}$ in the fitting constants method of estimation does not lead from (123) and (124) to (137) and (138) because, as Thompson (1969) points out, $R^*(\mathbf{b}, \mathbf{u})$ is not a reduction in the sum of squares due to solving (135).

It is true that

$$R^*(\mathbf{b}, \mathbf{u}) = \mathbf{b}^{*\prime}\mathbf{X}'\mathbf{y} + \mathbf{u}^{*\prime}\mathbf{Z}'\mathbf{y}.$$

However, the right-hand side of the equation is the reduction in the sum of squares only when the equation from which it stems, (135) in this case, is, for some matrix

$\mathbf{W}$, of the form

$$\mathbf{W}'\mathbf{W}\begin{bmatrix}\mathbf{b}^*\\ \mathbf{u}^*\end{bmatrix} = \mathbf{W}'\mathbf{y}.$$

By observation, (135) is not of this form. Furthermore, as shown by Thompson (1969), the reduction sum of squares after solving (135) is

$$\mathbf{y}'\mathbf{y} - (\mathbf{y} - \mathbf{X}\mathbf{b}^* - \mathbf{Z}\mathbf{u}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^* - \mathbf{Z}\mathbf{u}^*) = R^*(\mathbf{u}, \mathbf{b}) + \lambda\mathbf{u}^{*\prime}\mathbf{u}^*.$$

The calculations involved in the estimators (123) and (124) are summarized in Section 7b of Chapter 11 and those for the estimators (152) and (153) are in Section 7c of Chapter 11. Please see the web page or the first edition.

## 10. BEST QUADRATIC UNBIASED ESTIMATION

The variance component analogue of the best linear unbiased estimator (b.l.u.e.) of a function of fixed effects is a best quadratic unbiased estimator (BQUE) of a variance component. By this, we mean a quadratic function of the observations that is an unbiased estimator of the component and of all such estimators, it is the one with minimum variance. The BQUE's of variance components from balanced data are derived by the analysis of variance method as has been discussed in Section 8a of Chapter 9. As one might expect, derivation of such estimators from unbalanced data is more difficult. Ideally, we would like estimators that are uniformly "best" for all values of the variance components.

Townsend and Searle (1971) have obtained locally BQUE's for the variance components in a one-way classification model with $\mu = 0$ From these, they have suggested approximate BQUE's for the $\mu \neq 0$ model. The method used for the case where $\mu = 0$ is essentially the same as that used to derive the MIVQUE (see Section 10b(ii) of Chapter 9). Swallow and Searle (1978) use the MIVQUE method outlined in Section 10b(ii) of Chapter 9 to develop minimum variance quadratic unbiased estimators for the variance components for the case where $\mu \neq 0$. The case where $\mu = 0$ is included in their derivation. The method used by Townsend and Searle (1971) to find approximate BQUE's for $\mu \neq 0$ is different from that of Swallow and Searle. We first outline the development of BQUE's by Townsend and Searle (1971) for the case of a zero mean ($\mu = 0$). We then outline Swallow and Searle's (1978) development of the MIVQUE for the one-way classification model.

### a. The Method of Townsend and Searle (1971) for a Zero Mean

We write the model $y_{ij} = \alpha_i + e_{ij}$ similar to (94) as $\mathbf{y} = \mathbf{Z}\alpha + \mathbf{e}$ with $\mathbf{V}$ of (96) being $\mathbf{V} = \sigma_\alpha^2\mathbf{Z}\mathbf{Z}' + \sigma_e^2\mathbf{I}$. Suppose, we let the desired estimators of $\sigma_e^2$ and $\sigma_\alpha^2$ take the form

$$\sigma_e^2 = \mathbf{y}'\mathbf{A}\mathbf{y} \quad \text{and} \quad \sigma_\alpha^2 = \mathbf{y}'\mathbf{B}\mathbf{y},$$

together with the unbiasedness condition

$$E\left(\hat{\sigma}_e^2\right) = \text{tr}(\mathbf{AV}) = \sigma_e^2 \quad \text{and} \quad E\left(\hat{\sigma}_\alpha^2\right) = \text{tr}(\mathbf{BV}) = \sigma_\alpha^2 \tag{154}$$

such that

$$v\left(\hat{\sigma}_e^2\right) = 2\text{tr}(\mathbf{AV})^2 \quad \text{and} \quad v\left(\hat{\sigma}_\alpha^2\right) = 2\text{tr}(\mathbf{BV})^2 \text{ is minimized.} \tag{155}$$

The problem is then to find matrices $\mathbf{A}$ and $\mathbf{B}$ such that (155) is satisfied subject to (154). We obtain the canonical form of $\mathbf{V}$ under orthogonal similarity as $\mathbf{P'VP} = \mathbf{D}$, where $\mathbf{P}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix of eigenvalues (latent roots) of $\mathbf{V}$.

We then find that satisfying (154) and (155) demands minimizing $2\text{tr}(\mathbf{DQ})^2$ subject to $\sigma_e^2 = \text{tr}(\mathbf{DQ})$ and minimizing $2\text{tr}(\mathbf{DR})^2$ subject to $\sigma_\alpha^2 = \text{tr}(\mathbf{DR})$, where $\mathbf{Q} = \mathbf{P'AP}$ and $\mathbf{R} = \mathbf{P'BP}$. The eigenvalues of $\mathbf{V}$ are $\sigma_e^2$, with multiplicity $N - a$, and $\sigma_e^2 + n_i\sigma_\alpha^2$ for $i = 1, 2, \ldots, a$. The corresponding eigenvectors (latent vectors) are the columns of the matrix $\Sigma^+\mathbf{G}_i$, where $\mathbf{G}_i'$ is the last $(n_i - 1)$ rows of a Helmert matrix of order $n_i$ (see Section 1 of Chapter 2) and the columns of $\mathbf{Z}$**.** The minimization procedure leads, after some algebraic simplification, to the following results. Define

$$\rho = \frac{\sigma_\alpha^2}{\sigma_e^2}, \qquad r = \sum_{i=1}^{a} \frac{1}{(1 + n_i\rho)^2} + N - a,$$
$$s = \sum_{i=1}^{a} \frac{n_i^2}{(1 + n_i\rho)^2}, \quad \text{and} \quad t = \sum_{i=1}^{a} \frac{n_i}{(1 + n_i\rho)^2}.$$

Then, the BQUE's are

$$\hat{\sigma}_e^2 = \frac{1}{rs - t^2}\left[\sum_{i=1}^{a} \frac{s - tn_i}{(1 + n_i\rho)^2} \cdot \frac{y_{i.}^2}{n_i} + s(\text{SSE})\right] \tag{156a}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{rs - t^2}\left[\sum_{i=1}^{a} \frac{rn_i - t}{(1 + n_i\rho)^2} \cdot \frac{y_{i.}^2}{n_i} - t(\text{SSE})\right], \tag{156b}$$

where SSE is the usual error sum of squares, $\sum\sum y_{ij}^2 - \sum n_i\bar{y}_{i.}^2$.

These estimators are functions of the variance components because they are functions of the ratio $\rho = \sigma_\alpha^2/\sigma_e^2$. The variances of the estimators are identical to those of the large sample maximum likelihood estimators. The limits of the estimators as $\rho \to 0$ are the Koch (1968) estimators given in (92). The limit of $\hat{\sigma}_e^2$ as $\rho \to \infty$ is the analysis of variance method estimator of $\sigma_e^2$.

Two reasons why comparison of the BQUE's is difficult are:

(i) Their variances are functions of the unknown variance components;
(ii) The BQUE's themselves as in (156) are also functions of the unknown variance components.

Therefore, Townsend (1968) compared the BQUE's with the analysis of variance method (ANOVA) estimators numerically. In doing so, he used a range of values of $\rho$, both for the actual BQUE's and for approximate BQUE's using a prior estimate, or guess $\rho_0$ of $\rho$ in the estimation procedure. When the approximate BQUE is used in place of the ANOVA estimator, Townsend (1968) found that considerable reduction in the variance of $\sigma_\alpha^2$ can be achieved. In fact, this advantage can be gained even when rather inaccurate prior estimates (guesses) of $\rho$ are used as $\rho_0$. The reduction in variance appears to be greatest when the data are severely unbalanced and $\rho$ is either small or large. It appears smallest for values of $\rho$ that are moderately small. In some cases, there is actually no reduction in variance, when the ANOVA is a BQUE for some specific $\rho$. Details of these comparisons are available in Townsend (1968). The estimators, their variances and suggested expressions for the $\mu \neq 0$ model, taken from Townsend (1968) are available in Section 1f of Chapter 11 (see the web page or the first edition).

### b. The Method of Swallow and Searle (1978) for a Non-Zero Mean

Recall that the MIVQUE of

$$\mathbf{p}'\hat{\sigma} = \sum_{i=1}^{k} p_i \sigma_i^2$$

is taken to be a quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is a symmetric matrix chosen to minimize the variance of $\mathbf{y}'\mathbf{A}\mathbf{y}$ subject to $\mathbf{A}\mathbf{X} = \mathbf{0}$ and $\text{tr}(\mathbf{A}\mathbf{V}_i) = p_i$. C. R. Rao (1971b) shows that for

$$\mathbf{R} = \mathbf{V}^{-1}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}], \tag{157}$$

$$\mathbf{S} = \{s_{ii'}\} = \{\text{tr}(\mathbf{V}_i\mathbf{R}\mathbf{V}_{i'}\mathbf{R}\}, i, i' = 1, \ldots, k, \tag{158}$$

and

$$\mathbf{u} = \{u_i\} = \{\mathbf{y}'\mathbf{R}\mathbf{V}_i\mathbf{R}\mathbf{y}\}, i = 1, \ldots, k, \tag{159}$$

under normality, the vector of MIVQUE is

$$\hat{\sigma}^2 = \mathbf{s}^{-1}\mathbf{u}. \tag{160}$$

Under normality for balanced data for all of the usual models including fixed effects, mixed or random models for the one- and two-way classification with and without

interaction and nested models the MIVQUE estimators and the analysis of variance estimators are the same.

Using equation (10) in Chapter 9 for the one-way classification model, we have that the vector of observations $\mathbf{y}$ has a multivariate normal distribution with mean $\mathbf{1}_N\mu$ and variance–covariance matrix

$$\mathbf{V} = \sum{}^{+}\left(\sigma_e^2\mathbf{I}_{n_i} + \sigma_\alpha^2\mathbf{J}_{n_i}\right).$$

After simplifying the expressions (157)–(159), we write

$$k_i = \frac{n_i}{\sigma_e^2 + n_i\sigma_\alpha^2}, k = \frac{1}{\sum_{i=1}^{a} k_i} \text{ and } \bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}. \tag{161}$$

Then, the s's and the u's of (158) and (159) may be written

$$s_{11} = \sum_{i=1}^{a} k_i^2 - 2k\sum_{i=1}^{a} k_i^3 + k^2\left(\sum_{i=1}^{a} k_i^2\right)^2, \tag{162}$$

$$s_{12} = \sum_{i=1}^{a} \frac{k_i^2}{n_i} - 2k\sum_{i=1}^{a} \frac{k_i^3}{n_i} + k^2\sum_{i=1}^{a} k_i^2\sum_{i=1}^{a} \frac{k_i^2}{n_i}, \tag{163}$$

$$s_{22} = \frac{N-a}{\sigma_e^4} + \sum_{i=1}^{a} \frac{k_i^2}{n_i^2} - 2k\sum_{i=1}^{a} \frac{k_i^3}{n_i^2} + k^2\left(\sum_{i=1}^{a} \frac{k_i^2}{n_i}\right)^2, \tag{164}$$

$$u_1 = \sum_{i=1}^{a} k_i^2\left[\bar{y}_{i.} - k\sum_{i=1}^{a} k_i\bar{y}_{i.}\right]^2, \tag{165}$$

and

$$u_2 = \frac{1}{\sigma_e^4}\text{SSE} + \sum_{i=1}^{a} \frac{k_i^2}{n_i}\left[\bar{y}_{i.} - k\sum_{i=1}^{a} k_i\bar{y}_{i.}\right]^2. \tag{166}$$

From (160), under normality, the MIVQUE's of $\sigma_e^2$ and $\sigma_\alpha^2$ are then

$$\hat{\sigma}_e^2 = \frac{1}{|\mathbf{S}|}[-s_{12}u_1 + s_{11}u_2] \tag{167}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{|\mathbf{S}|}[s_{22}u_1 - s_{12}u_2], \tag{168}$$

where $|\mathbf{S}| = s_{11}s_{22} - s_{12}^2$.

The variances and covariances of these MIVQUE's are

$$v(\hat{\sigma}_e^2) = \frac{2s_{11}}{|\mathbf{S}|}, \tag{169}$$

$$v(\hat{\sigma}_\alpha^2) = \frac{2s_{22}}{|\mathbf{S}|}, \tag{170}$$

and

$$\operatorname{cov}(\hat{\sigma}_e^2, \hat{\sigma}_\alpha^2) = \frac{-2s_{12}}{|\mathbf{S}|}. \tag{171}$$

Unfortunately, the MIVQUE's are functions of the unknown variance components. Therefore, we must replace the unknown values of $\sigma_\alpha^2$ and $\sigma_e^2$ by some numbers $\sigma_{e0}^2$ and $\sigma_{\alpha 0}^2$ that are a priori estimates of $\sigma_\alpha^2$ and $\sigma_e^2$. Swallow and Searle (1978) give comparisons of the variances of the MIVQUE of $\sigma_\alpha^2$ for different a priori estimates and observe that in every case, the MIVQUE has a smaller variance than the analysis of variance method estimators.

In a practical problem, one way to choose a priori estimates is to consider the results of a previous run of the experiment or process, if available, and use the analysis of variance estimates obtained from the past data.

**Example 8 Numerical Comparison of Variances of Analysis of Variance Estimators and MIVQUE Estimators** The data for this example is taken from Swallow and Searle (1978).

Five groups of several consecutive bottles each were snatched from a moving production line that was filling the bottles with vegetable oil. The oil in each bottle was weighed. The data appears in the table below. A multiple (24) head machine was being used in the filling. Different (unidentified) heads are represented in the five groups of bottles sampled. Thus, in part, variability among groups reflects variability among heads.

| Net Weights (oz.) of Vegetable Oil Fills | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 15.70 | 15.69 | 15.75 | 15.68 | 15.65 |
| 15.68 | 15.71 | 15.82 | 15.66 | 15.60 |
| 15.64 | | 15.75 | 15.59 | |
| 15.60 | | 15.71 | | |
| | | 15.84 | | |

We have the Analysis of variance table.

**The SAS System**

**The GLM Procedure**

**Dependent Variable: weight**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 4 | 0.05530708 | 0.01382677 | 6.46 | 0.0063 |
| **Error** | 11 | 0.02353667 | 0.00213970 | | |
| **Corrected Total** | 15 | 0.07884375 | | | |

We also have the expected mean square of the model sum of squares.

**The SAS System**

**The GLM Procedure**

| Source | Type III Expected Mean Square |
|---|---|
| **group** | Var(Error) + 3.0938 Var(group) |

We have that

$$\begin{aligned} \hat{\sigma}_e^2 &= 0.0021397 \\ 3.0938\hat{\sigma}_\alpha^2 + 0.0021397 &= 0.01382677 \\ \hat{\sigma}_\alpha^2 &= 0.00378921 \end{aligned}$$

are the analysis of variance estimates for the variance components.

Suppose, from other data of the same type, we have analysis of variance estimates $\hat{\sigma}_\alpha^2 = 0.0028$ and $\hat{\sigma}_e^2 = 0.0025$. Using these estimates, Swallow and Searle find that the MIVQUE's are $\hat{\sigma}_\alpha^2 = 0.0021$ and $\hat{\sigma}_e^2 = 0.0032$. The MIVQUE have approximate estimated variances and covariances

$$v\left(\hat{\hat{\sigma}}_e^2\right) = 0.0000113, v\left(\hat{\hat{\sigma}}_\alpha^2\right) = 0.0000684, \text{ and } \operatorname{cov}\left(\hat{\sigma}_e^2, \hat{\hat{\sigma}}_\alpha^2\right) = -0.0000039.$$

The variances and covariances of the analysis of variance estimators are

$$v\left(\hat{\hat{\sigma}}_e^2\right) = 0.0000080, v\left(\hat{\hat{\sigma}}_\alpha^2\right) = 0.00001084, \text{ and } \operatorname{cov}\left(\hat{\sigma}_e^2, \hat{\hat{\sigma}}_\alpha^2\right) = -0.0000026.$$

Swallow (1981) compares the variances of MIVQUE's with the analysis of variance estimators. He notes from numerical comparisons that when $\sigma_\alpha^2/\sigma_e^2 > 1$ and unless $\sigma_{\alpha 0}^2/\sigma_{e0}^2 \leq \sigma_\alpha^2/\sigma_e^2$, where $\sigma_{\alpha 0}^2$ and $\sigma_{e0}^2$ are prior values of $\sigma_\alpha^2$ and $\sigma_e^2$:

(i) the MIVQUE's have variances near their lower bounds;
(ii) the MIVQUE of $\sigma_\alpha^2$ is more efficient than the analysis of variance estimator.

He also observes that:

(i) when $\sigma_\alpha^2/\sigma_e^2 < 1$, the MIVQUE's are more dependent on accurate specification of the ratio of the variance components $\sigma_{\alpha 0}^2/\sigma_{e0}^2$;

(ii) the MIVQUE and the analysis of variance estimator of $\sigma_e^2$ have nearly equal variances unless $\sigma_{\alpha 0}^2/\sigma_{e0}^2 \leq \sigma_\alpha^2/\sigma_e^2$ in which case the analysis of variance estimator has smaller variance.

By doing a Monte Carlo study, Swallow and Monahan (1984) compare the biases and mean square errors of the analysis of variance estimators, the MIVQUE, the restricted maximum likelihood estimator and maximum likelihood estimators of variance components in the one-way classification model. Their results indicate that:

(i) analysis of variance estimators perform well when $\sigma_\alpha^2/\sigma_e^2 > 1$;
(ii) when $\sigma_\alpha^2/\sigma_e^2 < 0.5$, maximum likelihood estimators are excellent;
(iii) MIVQUE with a priori estimators the analysis of variance estimators are adequate;
(iv) MIVQUE with a priori values $\sigma_{\alpha 0}^2 = 0$ and $\sigma_{e0}^2 = 1$ performs poorly when $\sigma_\alpha^2/\sigma_e^2 > 1$.

In our discussion of the MIVQUE, we have focused on the one-way classification model. P. S. R. S. Rao and Heckler (1997) consider the comparison of the variances and biases of analysis of variance, restricted maximum likelihood estimators and MIVQUE for a two-factor random-effects model with one factor nested.

## 11. SHRINKAGE ESTIMATION OF REGRESSION PARAMETERS AND VARIANCE COMPONENTS

We shall consider shrinkage estimators of regression parameters and of variance components. First, we explain what shrinkage estimators are and how they can be more efficient than maximum likelihood estimators. We then discuss the celebrated James–Stein estimator in the linear models context. Finally, we give examples of some improved estimators of the variance and of variance components.

### a. Shrinkage Estimators

Suppose we take one of the standard estimators of a vector of regression parameters or the variance–covariance matrix and multiply it by a constant between zero and one or a matrix $\mathbf{M}$, where $\mathbf{I} - \mathbf{M}$ is positive definite? The estimator we obtain in this way is called a shrinkage estimator. Usually shrinkage estimators, although biased, have a smaller variance than the estimators they shrink. In addition, in comparison to the standard estimators that are multiplied by the shrinkage factor they typically have a smaller mean square error for a range of the parameters. One example of such a shrinkage estimator is the ridge regression estimator of Hoerl and Kennard (1970) that was mentioned in Section 3 of Chapter 3. Notice that, for $k > 0$, we can write the

ridge estimator as

$$\begin{aligned}\mathbf{p}'\beta^{(r)} &= (\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{b} = (\mathbf{I}+k(\mathbf{X}'\mathbf{X})^{-1})^{-1}\mathbf{b}, \end{aligned} \tag{172}$$

where $\mathbf{b}$ is the least-square estimator. This is the product of a matrix whose difference between it and the identity matrix is positive definite and the least-square estimator.

### b. The James–Stein Estimator

Consider the linear Bayes estimator (see Section 3 of Chapter 3) derived from a prior distribution, where $\theta = 0$ and $\mathbf{F} = \frac{\sigma^2}{k}(\mathbf{X}'\mathbf{X})^{-1}$. The resulting estimator, as the reader may show in Exercise 15, is

$$\mathbf{p}'\hat{\beta}^{(c)} = \frac{1}{1+k}\mathbf{b} = \left(1 - \frac{k}{1+k}\right)\mathbf{b}. \tag{173}$$

The shrinkage estimator in (173) is the contraction estimator of Mayer and Willke (1973).

Dempster (1973) and Zellner (1986) suggested the prior distribution that resulted in the linear Bayes estimator of (173). Assuming a normal population, suppose we substitute the unbiased estimator $(s-2)\hat{\sigma}^2/\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$ for fraction $k/(1+k)$. Then, we have in place of (173),

$$\mathbf{p}'\beta^{(JS)} = \left(1 - \frac{(s-2)\hat{\sigma}^2}{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}\right)\mathbf{b}, \tag{174}$$

the celebrated James–Stein estimator in the context of a linear model. Gruber (1998) studies the James–Stein estimator for different linear model setups.

Stein (1956) and later James and Stein (1961) showed that the usual maximum likelihood estimator of the mean of a multivariate normal distribution is inadmissible. An estimator is inadmissible if we can find an estimator whose mean square error is strictly less than it for at least one point and less than or equal to it for the entire parameter space. An admissible estimator is one that we cannot do this for. We can show that the mean square error of (174) is less than that of the least-square estimator (see for example Gruber (1998)). The technique used to obtain (174) from (173), namely, replacing a function of the prior parameters by an estimator based on the data is known as empirical Bayes (see, for example, Efron and Morris (1973)).

### c. Stein's Estimator of the Variance

Stein (1964) showed that the minimum mean square error estimator of the variance of a normal population is inadmissible. The inadmissible estimator is

$$\hat{\sigma}^2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{175}$$

Stein shows that the estimator (175) is inadmissible by establishing that

$$\hat{\sigma}_1^2 = \min\left\{\frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \frac{1}{n+2}\sum_{i=1}^{n}(x_i-\varsigma_0)^2\right\}, \tag{176}$$

for any fixed number $\varsigma_0$, has a mean square error strictly smaller than that of (175) for at least one point of the parameter space and less than or equal to that of (175) for the entire parameter space. Both the estimators in (174) and (176) are also inadmissible. We can construct an estimator with smaller mean square error than (174) by truncating the shrinkage factor to be zero for values where it would be negative. Brewster and Zidek (1974), and Brown (1968) also produce estimators with smaller mean square error than (175).

### d. A Shrinkage Estimator of Variance Components

We shall present an estimator of the variance components for a mixed linear model that has a smaller mean square error than that of the analysis of variance estimator. The estimator is due to Kubokawa (1995). We shall follow this paper in our presentation.

We consider a general linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \mathbf{e}, \tag{177}$$

where $\mathbf{y}$ is an $n$-vector of observations, $\mathbf{X}$ is an $n \times p_1$ known matrix with rank$(\mathbf{X}) = r$ and $\beta$ is a $p_1$-vector of parameters. In addition, $\mathbf{Z}$ is a given $n \times p_2$ matrix, $\alpha$ and $\mathbf{e}$ are independent random $p_2$- and $n$-vectors, respectively with $\alpha \sim N_{p_2}(\mathbf{0}, \sigma_A^2\mathbf{I})$ and $\mathbf{e} \sim N_n(\mathbf{0}, \sigma_e^2\mathbf{I}_n)$. The random one-way analysis of variance model is a special case.

Consider an $(n-r)\times n$ matrix $\mathbf{P}_1$ and an $r \times n$ matrix $\mathbf{P}_2$ such that $\mathbf{P}_1\mathbf{X} = \mathbf{0}, \mathbf{P}_1\mathbf{P}_2' = \mathbf{0}, \mathbf{P}_1\mathbf{P}_1' = \mathbf{I}_{n-r}$, and $\mathbf{P}_2\mathbf{P}_2' = \mathbf{I}_r$. Such matrices exist by the singular value decomposition. Let $\mathbf{x}_1 = \mathbf{P}_1\mathbf{y}$ and $\mathbf{x}_2 = \mathbf{P}_2\mathbf{y}$. It follows that

$$\mathbf{x}_1 \sim N_{n-r}\left(\mathbf{0}, \sigma_A^2\mathbf{P}_1\mathbf{Z}\mathbf{Z}'\mathbf{P}_1' + \sigma_e^2\mathbf{I}_{n-r}\right)$$

and

$$\mathbf{x}_2 \sim N_r\left(\mathbf{P}_2\mathbf{X}\beta, \sigma_A^2\mathbf{P}_2\mathbf{Z}\mathbf{Z}'\mathbf{P}_2' + \sigma^2\mathbf{I}_r\right).$$

Consider the spectral decompositions

$$\mathbf{P}_1\mathbf{Z}\mathbf{Z}'\mathbf{P}_1' = \sum_{i=1}^{l}\lambda_i\mathbf{E}_{1i} \quad \text{and} \quad \mathbf{P}_2\mathbf{Z}\mathbf{Z}'\mathbf{P}_2' = \sum_{j=1}^{k-1}\tau_j\mathbf{E}_{2j},$$

where $\text{rank}(\mathbf{E}_{1i}) = m_i$, $\sum_{i=1}^{l} m_i = \text{rank}(\mathbf{P}_1\mathbf{Z}\mathbf{Z}'\mathbf{P}_1')$, $\text{rank}(\mathbf{E}_{2j}) = n_j$, and $\sum_{j=1}^{k-1} n_j = \text{rank}(\mathbf{P}_2\mathbf{Z}\mathbf{Z}'\mathbf{P}_2')$. Assume that $\lambda_i > 0$ and $\tau_j > 0$ satisfy

$$0 < \lambda_1 < \cdots < \lambda_l \text{ and } \tau_1 > \cdots \tau_{k-1} > 0. \tag{178}$$

Assume that

$$m = n - r - \sum_{i=1}^{l} m_i > 0. \tag{179}$$

Let $E_{1,l+1} = I_{n-r} - \sum_{i=1}^{l} E_{1i}$ and $E_{2,k} = I_r - \sum_{j=1}^{k-1} E_{2j}$. We see that $\text{rank}(E_{1,i+1}) = m > 0$ and $\text{rank}(E_{2,k}) = r - \sum_{j=1}^{k-1} n_j = n_k \geq 0$. Thus, we obtain the quadratic statistics

$$\begin{aligned} S &= x_1' E_{1,i+1} x_1 \sim \sigma^2 \chi^2_m, \\ S_1 &= x_1' E_{11} x_1 \sim \left(\sigma^2 + \lambda_1 \sigma_A^2\right) \chi^2_{m_1}, \\ &\cdots \\ S_l &= x_1' E_{1l} x_1 \sim \left(\sigma^2 + \lambda_l \sigma_A^2\right) \chi^2_{m_l} \end{aligned} \tag{180a}$$

and

$$\begin{aligned} T_1 &= x_2' E_{21} x_2 \sim \left(\sigma^2 + \tau_1 \sigma_A^2\right) \chi^2 \left( n_1, \frac{\beta' \mathbf{X}' \mathbf{P}_2' \mathbf{E}_{2,i} \mathbf{P}_2 \mathbf{X} \beta}{2\left(\sigma^2 + \tau_1 \sigma_A^2\right)} \right) \\ &\cdots \\ T_k &= x_2' E_{2,k} x_2 \sim \left(\sigma^2 + \tau_k \sigma_A^2\right) \chi^2 \left( n_k, \frac{\beta' \mathbf{X}' \mathbf{P}_2' \mathbf{E}_{2,k} \mathbf{P}_2 \mathbf{X} \beta}{2\left(\sigma^2 + \tau_k \sigma_A^2\right)} \right). \end{aligned} \tag{180b}$$

The analysis of variance estimator that one derives by Henderson's method 3 is

$$\hat{\sigma}_A^2 = \frac{\sum_{i=1}^{l} m_i}{\sum_{i=1}^{l} \lambda_i m_i} \left( \frac{\sum_{i=1}^{l} S_i}{\sum_{i=1}^{l} m_i} - \frac{S}{m} \right). \tag{181}$$

Kubokawa (1995) shows that an estimator with a smaller mean square error than that of (181) takes the form

$$\hat{\sigma}_{0A}^2 = \frac{\sum_{i=1}^{l} m_i}{\sum_{i=1}^{l} \lambda_i m_i} \left( \frac{\sum_{i=1}^{l} S_i}{\sum_{i=1}^{l} m_i + 2} - \frac{S}{m} \right) \tag{182}$$

and that an estimator with smaller mean square error than (182) takes the form

$$\hat{\sigma}^2_{1A} = \max\left\{\hat{\sigma}^2_{1A}, \frac{2\sum_{i=1}^{l} m_i}{m\sum_{i=1}^{l}\lambda_i m_i\left(\sum_{i=1}^{l} m_i + 2\right)} S\right\} \tag{183}$$

More discussion of shrinkage estimators for variance components is available in Kubokawa (1999), Cui et al. (2005), An (2007), and the references therein.

## 12. EXERCISES

**1** In Example 8, use the analysis of variance estimate for $\hat{\sigma}^2_{\alpha}$ and $\hat{\sigma}^2_{e}$ as prior estimates to find the MIVQUE's. How does the variance of these estimates compare with the analysis of variance estimates?

**2** In the $r$-way classification random model, having all possible interactions, show that var(**t**) has $2^{r-1}(2^r + 1)$ different elements.

**3** For balanced data, show that:

**(a)** In equation (32), $\theta_1 = bn\left(\sum_{i=1}^{a}\alpha_i^2 - a\bar{\alpha}_{\cdot}^2\right)$.

**(b)** For random $\alpha$, thatis,$\alpha_i \sim N(0, \sigma^2_{\alpha}), E(\theta_1) = (a-1)bn\sigma^2_{\alpha}$.

**4** Establish the following for result (33)

**(a)** For random $\alpha$'s

$$E(\theta_2) = \left(\sum_{j=1}^{b}\frac{\sum_{i=1}^{a} n_{ij}^2}{n_{\cdot j}} - \frac{\sum_{i=1}^{a} n_{i\cdot}^2}{N}\right)\sigma^2_{\alpha}$$

**(b)** For balanced data, $\theta_2 = 0$.

**5** Consider the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where $x_i \sim N(\mu, \sigma^2)$. Assume that the $x_i$'s are independent.

**(a)** After making the transformation $y_i = x_i - \mu$, show that

$$s^2 = \frac{1}{n-1}\mathbf{y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{y}$$

**(b)** Using Corollaries 4.1 and 4.3 of Chapter 2, show that

$$E(s^2) = \sigma^2$$

and

$$\text{var}(s^2) = \frac{2\sigma^4}{n-1}$$

**6** Show that $\hat{\mathbf{v}} = (\mathbf{I} + \mathbf{A})^{-1}\mathbf{A}\hat{\gamma}$ is an unbiased estimator of $\mathbf{v}$ (see equations (56)–(58)).

**7** Show that the estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_\alpha^2$ given by (87)–(89) are the analysis of variance estimators for balanced data.

**8** Find the variance of $\hat{\sigma}_\alpha^2$ that can be derived from (88) and (89).

**9** Show that for balanced data, the estimator in (90) simplifies to the analysis of variance estimator.

**10** **(a)** Check the term below (91) for finding the variance of $\hat{\sigma}_e^2$ given in (90).

**(b)** Find the variance of $\hat{\sigma}_e^2$ given in (90).

**(c)** What does this variance simplify to, in the balanced case?

**11** Consider the one-way random model

$$\begin{array}{l} y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \ldots, a, j = 1, 2, \ldots, n_i \\ \alpha_i \sim N\left(0, \sigma_\alpha^2\right), e_{ij} \sim N(0, \sigma_e^2) \end{array}.$$

Define

$$T_0 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 \quad \text{and} \quad T_\mu = \frac{y_{..}}{N}.$$

Show from first principles that

**(a)** $E(T_0) = N(\mu^2 + \sigma_\alpha^2 + \sigma_e^2)$

**(b)** $E(T_\mu) = N\mu^2 + \sum_{i=1}^{a}\frac{n_i^2}{N} + \sigma_e^2$

**12** Show that for a balanced one-way ANOVA model, the MIVQUE of the variance components are those obtained by the ANOVA method.

**13** In fitting $\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_f\mathbf{b}_f + \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$,

**(a)** show that $R(\mathbf{b}_1|\mathbf{b}_f)$ equals $R(\mathbf{b}_1)_z$ when fitting $\mathbf{z} = \mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{X}_1\mathbf{b}_1 + \mathbf{W}\mathbf{e}$, where

$$\mathbf{W} = \mathbf{I} - \mathbf{X}_f(\mathbf{X}_f'\mathbf{X}_f)^{-1}\mathbf{X}_f'.$$

**(b)** Show that the reduction in the sum of squares due to fitting $\mathbf{z} = \mathbf{W}\mathbf{X}_1\mathbf{b}_1 + \mathbf{W}\mathbf{e}$ is $R(\mathbf{b}_1)_z$.

**14** **(a)** Show that the generalized inverse in (133) is indeed a generalized inverse by direct computation.

**(b)** Use the result of (a) to establish (129).

**(c)** Derive the last equation at the end of Section 9.

**15** **(a)** Show how to derive (173) as a linear Bayes estimator from the given prior assumptions.

**(b)** Find a range of values of the $\beta$ parameters where (173) has a smaller mean square error than the least-square estimator.

# REFERENCES


Adke, S. R. (1986). One-way ANOVA for dependent observations. *Commun. Stat. Theory Methods*, **15**, 1515–1528.

Ahrens, H. (1965). Standardfehler geschatzter varianzkomponenten eines unbalanzie versuchplanes in *r*-stufiger hierarchischer klassifikation. *Monatsb. Deut. Akad. Wiss. Berlin*, **7**, 89–94.

Al Sarraj, R., and D. von Rosen. (2008). Improving Henderson's method 3 approach when estimating variance components in a two-way mixed model. In: *Statistical Inference Econometric Analysis and Matrix Algebra*, B. Schipp and W. Kräer (Eds.), pp. 125–142. Physica-Verlag, Heidelberg.

An, L. (2007). Shrinkage estimation of variance components with applications to microarray data. Ph.D. Thesis, University of Windsor.

Anderson, R. D. (1978). Studies on the estimation of variance components. Ph.D. Thesis, Cornell University, Ithaca, NY.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edition. John Wiley & Sons, New York.

Anderson, R. L., and T. A. Bancroft. (1952). *Statistical Theory in Research*. McGraw-Hill, New York.

Anderson, R. L., and P. P. Crump. (1967). Comparison of designs and estimation procedures for estimating parameters in a two stage nested process. *Technometrics*, **9**, 499–516.

Andrews, D. F., and A. M. Herzberg. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York.

Anscombe, F. J., and J. W. Tukey. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141–160.



*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

Banerjee, K. S. (1964). A note on idempotent matrices. *Ann. Math. Stat.*, **35**, 880–882.

Bartlett, M. S. (1937). Some examples of statistical methods in research in agriculture and applied biology. *J. R. Stat. Soc. Suppl.*, **4**, 137–183.

Bartlett, M. S., and D. G. Kendall. (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *J. R. Stat. Soc. B*, **8**, 128–138.

Ben-Israel, A., and T. N. E. Greville. (2003). *Generalized Inverses Theory and Applications*, 2nd edition. Springer-Verlag, New York.

Bennett, C. A., and N. L. Franklin. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*. John Wiley & Sons, New York.

Blischke, W. R. (1966). Variances of estimates of variance components in a three-way classification. *Biometrics*, **22**, 553–565.

Blischke, W. R. (1968). Variances of moment estimators of variance components in the unbalanced *r*-way classification. *Biometrics*, **24**, 527–540.

Boullion, T. L., and P. L. Odell (Eds.). (1968). *Theory and Application of Generalized Inverse Matrices*. Mathematics Series No. 4, Texas Technical College, Lubbock, TX.

Brewster, J. F., and J. V. Zidek. (1974). Improving on equivariant estimators. *Ann. Stat.*, **2**(1), 21–38.

Broemeling, L. D. (1969). Confidence regions for variance ratios of random models. *J. Am. Stat. Assoc.*, **64**, 660–664.

Brorsson, B., J. Ifver, and H. Rydgren. (1988). Injuries from single vehicle crashes and snow depth. *Accid. Anal. Prev.*, **19**, 367–377.

Brown, K. G. (1977). On estimation of a diagonal covariance matrix by MINQUE. *Commun. Stat. A*, **6**(5), 471–484.

Brown, K. H. (1968). Social class and family investment. Ph.D. Thesis, Cornell University, Ithaca, NY.

Brown, L. (1968). Inadmissibility of the usual estimators of the scale parameter in problems with unknown location and scale parameters. *Ann. Math. Stat.*, **39**, 29–48.

Brown, M. B., and A. B. Forsyth. (1974a). Robust tests for the equality of variances. *J. Am. Stat. Assoc.*, **69**, 364–367.

Brown, M. B., and A. B. Forsyth. (1974b). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, **16**(1), 129–132.

Bulmer, M. G. (1957). Approximate confidence intervals for components of variance. *Biometrika*, **44**, 159–167.

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

Bush, N., and R. L. Anderson. (1963). A comparison of three different procedures for estimating variance components. *Technometrics*, **5**, 421–440.

Cameron, E., and L. Pauling. (1976). Supplemental ascorbate in the supportive treatment of cancer: prolongation of survival times in human terminal cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **73**, 3685–3689.

Casella, G., and R. L Berger. (2002). *Statistical Inference*, 2nd edition. Duxbury, Pacific Grove, CA.

Chaubey, Y. B. (1984). On the comparison of some non-negative estimators of variance components for two models. *Commun. Stat. B Simul. Comp.*, **13**, 619–633.

Chen, C. J., K. Hiramatsu, T. Ooue, K. Takagi, and T. Yamamoto. (1991). Measurement of noise-evoked blood pressure by means of averaging method: relation between blood pressure rise and SPL. *J. Sound Vib.*, **151**, 383–394.

Chipman, J. S. (1964). On least squares with insufficient observations. *J. Am. Stat. Assoc.*, **69**, 1078–1111.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system with applications to the analysis of variance. *Math. Proc. Camb. Philos. Soc.*, **30**, 178–191.

Cohen, A. (1991). Dummy variables in stepwise regression. *Am. Stat.*, **45**, 226–228.

Conover, W. J. (1998). *Practical Non-parametric Statistics*, 3rd edition. John Wiley & Sons, New York.

Conover, W. J., and R. L. Iman. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.*, **55**, 124–129.

Corbeil, R. R., and S. R. Searle. (1976). A comparison of variance component estimators. *Biometrics*, **32**, 779–791.

Cornfield, J., and J. W. Tukey. (1956). Average values of mean squares as factorials. *Ann. Math. Stat.*, **27**, 907–949.

Courant, R. (1988). *Differential and Integral Calculus*, vol. 2. John Wiley & Sons, New York.

Crump, S. L. (1947). The estimation of components of variance in multiple classifications. Ph.D. Thesis, Iowa State College Library, Ames, IA.

Cui, X., J. T. G Hwang, J. Qui, N. Blades, and G. A. Churchill. (2005). Improved statistical tests for gene expression by shrinking variance components. *Biostatistics*, **6**(1), 59–75.

Cunningham, E. P. (1969). A note on the estimation of variance components by the method of fitting constants. *Biometrika*, **56**, 683–684.

Cunningham, E. P., and C. R. Henderson. (1968). An iterative procedure for estimating fixed effects and variance components in mixed model situations. *Biometrics*, **24**, 13–25. Corrigenda: *Biometrics*, **25**, 777–778 (1969).

Daniel, W. W. (1990). *Applied Nonparametric Statistics*. PWS-KENT Publishing Company, Boston, MA.

Dempster, A. P. (1973). Alternatives to least squares in multiple regression. In: *Multivariate Statistical Inference*, R. P. Kabe and D. G. Gupta (Eds.). North Holland, Amsterdam.

Draper, N. R. (2002). Applied regression analysis bibliography update 2000–2001. *Commun. Stat. Theory Methods*, **31**, 2051–2075.

Draper, N. R., and H. Smith. (1998). *Applied Regression Analysis*, 3rd edition. John Wiley & Sons, New York.

Dudewicz, E. J., and T. A. Bishop. (1981). Analysis of variance with unequal variances. *J. Qual. Technol.*, **13**(2), 111–114.

Eccleston, J., and K. Russell. (1975). Connectedness and orthogonality in multi-factor designs. *Biometrika*, **62**(2), 341–345.

Efron, B., and C. Morris. (1973). Stein's estimation rule and its competitors. *J. Am. Stat. Assoc.*, **65**, 117–130.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, **20**, 681–698.

Elston, R. C., and N. Bush. (1964). The hypotheses that can be tested when there are interactions in an analysis of variance model. *Biometrics*, **20**, 681–698.

Evans, D. A. (1969). Personal communication with S. R. Searle.

Federer, W. T. (1955). *Experimental Design*. Macmillan, New York.

Gaylor, D. W., and T. D. Hartwell. (1969). Expected mean squares for nested classifications. *Biometrics*, **25**, 427–430.

Gaylor, D. W., and F. N. Hopper. (1969). Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics*, **11**, 691–706.

Gaylor, D. W., H. L., Lucas, and R. L. Anderson. (1970). Calculation of expected mean squares by the abbreviated Doolittle and square root methods. *Biometrics*, **26**, 641–656.

Giesbrecht, F. G. (1983). An efficient procedure for computing MINQUE of variance components and generalized least square estimators of fixed effects. *Comm. Stat. A*, **12**, 2169–2177.

Godolphin, J. D. (2013). On the connectivity problem for m-way designs. *J. Stat. Theory Pract.*, **7**(4), 732–744.

Goldman, A. J., and M. Zelen. (1964). Weak generalized inverses and minimum variance unbiased estimators. *J. Res. Natl. Bur. Stand.*, **68B**, 151–172.

Golub, G. H., and C. F. Van Loan. (1996). *Matrix Computations*, 3rd edition. Johns Hopkins University Press, Baltimore, MD.

Goslee, D. G., and H. L. Lucas. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometrics*, **21**, 115–133.

Graybill, F. A. (1954). On quadratic estimates of variance components. *Ann. Math. Stat.*, **25**, 367–372.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*, vol. 1. McGraw-Hill, New York.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Wadsworth, Boston, MA.

Graybill, F. A., and R. A. Hultquist. (1961). Theorems concerning Eisenhart's Model II. *Ann. Math. Stat.*, **32**, 261–269.

Graybill, F. A., C. D. Meyer, and R. J. Painter. (1966). Note on the computation of the generalized inverse of a matrix. *SIAM Rev.*, **8**, 522–524.

Graybill, F. A., and G. Marsaglia. (1957). Idempotent matrices and quadratic forms in the general linear hypothesis. *Ann. Math. Stat.*, **28**, 678–686.

Graybill, F. A., and A. W. Wortham. (1956). A note on uniformly best unbiased estimators for variance components. *J. Am. Stat. Assoc.*, **51**, 266–268.

Greville, T. N. E. (1966). Note on the generalized inverse of a matrix product. *SIAM Rev.*, **8**, 518–521. Erratum: *SIAM Rev.*, **9**, 249 (1967).

Grubbs, F. E. (1948). On estimating precision of measuring instruments and product variability. *J. Am. Stat. Assoc.*, **43**, 243–264.

Gruber, M. H. J. (1998). *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Marcel Dekker, New York.

Gruber, M. H. J. (2010). *Regression Estimators: A Comparative Study*, 2nd edition. Johns Hopkins University Press, Baltimore, MD.

Gruber, M. H. J. (2014). *Matrix Algebra for Linear Models*. John Wiley & Sons, Hoboken, NJ.

Hartley, H. O. (1967). Expectation, variances and covariances of ANOVA mean squares by 'synthesis'. *Biometrics*, **23**, 105–114. Corrigenda: p. 853.

Hartley, H. O., and J. N. K. Rao. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Hartley, H. O., and S. R. Searle. (1969). On interaction variance components in mixed models. *Biometrics*, **25**, 573–576.

Harville, D. A. (1967). Estimability of variance components for the 2-way classification with interaction. *Ann. Math. Stat.*, **38**, 1508–1519.

Harville, D. A. (1969a). Quadratic unbiased estimation of variance components for the one-way classification. *Biometrika*, **56**, 313–326.

Harville, D. A. (1969b). Variance component estimation for the unbalanced one-way random classification—a critique. Aerospace Research Laboratories, ARL-69-0180, Wright Patterson Air Force Base, Ohio.

Harville, D. A. (1969c). Variances of the variance-component estimators for the unbalanced 2-way cross-classification with application to balanced incomplete block designs. *Ann. Math. Stat.*, **40**, 408–416.

Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.*, **72**, 320–340.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Henderson, C. R. (1959, 1969). Design and analysis of animal husbandry experiments. In: *Techniques and Procedures in Animal Science Research* (1st edition in 1959; 2nd edition in 1969), American Society of Animal Science.

Henderson, C. R. (1963). Selection index and expected genetic advance. In: *Statistical Genetics in Plant Breeding*. National Academy of Sciences, National Research Council Publication No. 982.

Henderson, C. R. (1968). Personal communication with S. R. Searle.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–448.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. N. Von Krosigk. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Herbach, L. H. (1959). Properties of type II analysis of variance tests. *Ann. Math. Stat.*, **30**, 939–959.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Am. Stat. Assoc.*, **60**, 806–825.

Hill, B. M. (1967). Correlated errors in the random model. *J. Am. Stat. Assoc.*, **62**, 1387–1400.

Hirotsu, C. (1966). Estimating variance components in a two-way layout with unequal numbers of observations. *Rept. Stat. Appl. Res., Union of Japanese Scientists and Engineers,* **13**(2), 29–34.

Hochberg, Y., and A. C. Tamhane. (1987). *Multiple Comparison Procedures*. John Wiley & Sons.

Hocking, R. R., and F. M. Speed. (1975). A full rank analysis of some linear model problems. *J. Am. Stat. Assoc.*, **70**, 706–712.

Hoerl, A. E., and R. W. Kennard. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hogg, R. V., J. W. Mc Kean, and A. T Craig. (2014). *Introduction to Mathematical Statistics*. Pearson Prentice Hall, Upper Saddle River, NJ.

Houf, R. E., and D. B. Berman. (1988). Statistical analysis of power module thermal test equipment. *IEEE Trans. Compon. Hybrids Manuf. Technol.*, **11**(4), 516–520.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, Boca Raton, FL.

James, W., and C. Stein. (1961). Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematics Statistics and Probability*, pp. 361–379. University of California Press, Berkeley, CA.

John, P. W. M. (1964). Pseudo-inverses in analysis of variance. *Ann. Math. Stat.*, **35**, 895–896.

Kaplan, J. S. (1983). A method for calculating MIVQUE estimators of variance components. *J. Am. Stat. Assoc.*, **78**, 476–477.

Karathanasis, A. D., and J. R. V. Pils. (2005). Solid-phase chemical fractionation of selected trace metals in some northern Kentucky soils. *Soil Sediment Contam. Int. J.*, **14**(4), 293–308.

Kelly, R. J., and T. Mathew. (1994). Improved nonnegative estimation of variance components in some mixed models with unbalanced data. *Technometrics*, **36**(2), 171–181.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* John Wiley & Sons, New York.

Kempthorne, O. (1968). Discussion of Searle [1968]. *Biometrics*, **24**, 782–784.

Kirk, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*. Brooks/Cole, Belmont, CA.

Kleffee, J., and B. Seifert. (1984). Matrix free computation of C. R. Rao's MINQUE for unbalanced nested classification models. *Comp. Stat. Data Anal.*, **2**(3), 692–698.

Koch, G. G. (1967a). A general approach to the estimation of variance components. *Technometrics*, **9**, 93–118.

Koch, G. G. (1967b). A procedure to estimate the population mean in random-effects models. *Technometrics*, **9**, 577–586.

Koch, G. G. (1968). Some further remarks on "A general approach to the estimation of variance components." *Technometrics*, **10**, 551–558.

Kubokawa, T. (1995). Estimation of variance components in mixed linear models. *J. Multivar. Anal.*, **53**, 210–236.

Kubokawa, T. (1999). Shrinkage and modification techniques in estimation of variance and related problems: a review. *Commun. Stat. Theory Methods*, **28**(3&4), 613–650.

Kussmaul, K., and R. L. Anderson. (1967). Estimation of variance components in two stage nested designs with complete samples. *Technometrics*, **9**, 373–390.

LaMotte, L. R., and R. R. Hocking. (1970). Computational efficiency in the selection of regression variables. *Technometrics*, **12**, 83–94.

Lancaster, H. O. (1954). Traces and cumulants of quadratic forms in normal variables. *J. R. Stat. Soc. B*, **16**, 247–254.

Lancaster, H. O. (1965). The Helmert matrices. *Am. Math. Mon.*, **72**, 4–12.

Lee, S., and C. H. Ahn. (2003). Modified ANOVA for unequal variances. *Comm. Stat. Simul. Comput.*, **32**(4), 987–1004.

Lehmann, E. L., and G. Casella. (1998). *Theory of Point Estimation*. Springer, New York.

Lehmann, E. L., and J. P. Romano. (2005). *Testing Statistical Hypotheses*, 3rd edition. Springer, New York.

Leone, F. C., L. S. Nelson., N. L. Johnson, and S. Eisenstat. (1968). Sampling distributions of variance components. II. Empirical studies of unbalanced nested designs. *Technometrics*, **10**, 719–738.

Levene, H. (1960). Robust tests for equality of variances. In: *Contributions to Probability and Statistics*, I. Olkin (Ed.), pp. 278–292. Stanford University Press, Palo Alto, CA.

Li, C. C. (1964). *Introduction to Experimental Statistics*. McGraw-Hill, New York.

Lou, L. M., and J. Senturia. (1977). Computation of MINQUE variance coefficient estimates. *J. Am. Stat. Assoc.*, **72**, 867–868.

Low, L. Y. (1964). Sampling variances of estimates of components of variance from a non-orthogonal two-way classification. *Biometrika*, **51**, 491–494.

Loynes, R. M. (1966). On idempotent matrices. *Ann. Math. Stat.*, **37**, 295–296.

Lum, M. D. (1954). Rules for determining error terms in hierarchal and partially hierarchal models. Wright Air Development Center, Dayton, OH.

Mahamunulu, D. M. (1963). Sampling variances of the estimates of variance components in the unbalanced 3-way nested classification. *Ann. Math. Stat*., **34**, 521–527.

Mannheim, B., and A. Cohen. (1978). Multivariate analysis of factors affecting work role centrality of occupational categories. *Hum. Relat.*, **31**, 525–553.

Mayer, L. W., and T. A. Willke. (1973). On biased estimation in linear models. *Technometrics*, **15**, 497–508.

Mc Elroy, F. W. (1967). A necessary and sufficient condition that ordinary least square estimators be best linear unbiased. *J. Am. Stat. Assoc.*, **62**, 1302–1304.

Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd edition. Springer-Verlag, Berlin and Heidelberg.

Miller, I., and M. Miller. (2012). *John Freund's Mathematical Statistics with Applications*, 6th edition. Pearson Prentice Hall, Upper Saddle River, NJ.

Millman, J., and G. V. Glass. (1967). Rules of thumb for writing the ANOVA table. *J. Educ. Meas.*, **4**, 41–51.

Mitra, S. K. (1971). Another look at Rao's MINQUE for variance components. *Bull. Int. Stat. Inst.*, **44**, 279–283.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*, 6th edition. John Wiley & Sons, Hoboken, NJ.

Montgomery, D. C., and G. C. Runger. (2007). *Applied Statistics and Probability for Engineers*, 4th edition. John Wiley & Sons, Hoboken NJ.

Mood, A. M. (1950). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

Mood, A. M., and F. A. Graybill. (1963). *Introduction to the Theory of Statistics*, 2nd edition. McGraw-Hill, New York.

Mood, A. M., F. A. Graybill, and C. Boes. (1974). *Introduction to the Theory of Statistics*, 3rd edition. McGraw-Hill, New York.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, **26**, 394–395.

Navidi, W. (2011). *Statistics for Engineers and Scientists*, 3rd edition. McGraw-Hill, New York.

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika*, **41**, 544–548.

Park, D. R., and K. R. Shah. (1995). On connectedness of row column designs. *Commun. Stat. Theory Methods*, **24**(1), 87–96.

Patterson, H. D., and R. Thompson. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Penrose, R. A. (1955). A generalized inverse for matrices. *Proc. Camb. Philos. Soc.*, **51**, 406–413.

Plackett, R. L. (1960). *Principles of Regression Analysis.* Oxford University Press.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

Raghavarao, D., and W. T. Federer. (1975). On connectedness in two-way elimination of heterogeneity designs. *Ann. Stat.*, **3**(3), 730–735.

Rao, C. R. (1962). A note on a generalized inverse of a matrix with application to problems in mathematical statistics. *J. R. Stat. Soc. B*, **24**, 152–158.

Rao, C. R. (1970). Estimation of heteroscedatic variances in linear models. *J. Am. Stat. Assoc.*, **65**, 161–172.

Rao, C. R. (1971a). Estimation of variance and covariance components—MINQUE theory. *J. Multivar. Anal.*, **1**, 257–275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *J. Multivar. Anal.*, **1**, 445–456.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *J. Am. Stat. Assoc.*, **67**, 112–115.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* John Wiley & Sons, London.

Rao, C. R. (1984). Optimization of functions of matrices with application to statistical problems. In: *W. G. Cochran's Impact on Statistics*, P. S. R. S. Rao and J. Sedransk (Eds.), pp. 191–202. John Wiley & Sons, New York.

Rao, C. R., and S. K. Mitra. (1971). *Generalized Inverses of Matrices and Its Applications*. John Wiley & Sons, New York.

Rao, C. R., and H. Toutenburg. (1999). *Linear Models: Least Squares and Alternatives*, 2nd edition. Springer, New York.

Rao, J. N. K. (1968). On expectations, variances and covariances of ANOVA mean squares by 'synthesis'. *Biometrics*, **24**, 963–978.

Rao, P. S. R. S. (1997). *Variance Component Estimation*. Chapman & Hall, UK.

Rao, P. S. R. S., and Y. P. Chaubey. (1978). Three modifications of the principle of the MINQUE. *Commun. Stat.*, **A7**, 767–778.

Rao, P. S. R. S., and C. Heckler. (1997). Estimators for the three-fold nested random effects model. *J. Stat. Plan. Inference*, **64**(2), 341–352.

Rao, P. S. R. S., J. Kaplan, and W. G. Cochran. (1981). Estimators for the one way random effects with unequal variances. *J. Am. Stat. Assoc.*, **76**, 89–97.

Rayner, A. A., and P. M. Pringle. (1967). A note on generalized inverses in the linear hypothesis not of full rank. *Ann. Math. Stat.*, **38**, 271–273.

Rice, W. R., and S. Gaines. (1989). Analysis of variance with unequal variances. *Proc. Natl. Acad. Sci. U. S. A.*, **86**(21), 8183–8184.

Robinson, J. (1965). The distribution of a general quadratic form in normal random variables. *Aust. J. Stat.*, **7**, 110–114.

Robinson, G. K. (1991). That BLUP is a good thing: estimation of random effects. *Stat. Sci.*, **6**(1), 15–51.

Robson, D. S. (1957). Personal communication with S. R. Searle.

Rohde, C. A. (1964). Contributions to the theory, computations and applications of generalized inverses. Institute of Statistics, Mimeo Series No. 392, University of North Carolina at Raleigh.

Rohde, C. A. (1965). Generalized inverses of partitioned matrices. *J. Soc. Ind. App. Math.*, **13**, 1033–1035.

Rohde, C. A. (1966). Some results on generalized inverses. *SIAM Rev.*, **8**, 201–205.

Rohde, C. A., and G. M. Tallis. (1969). Exact first- and second-order moments of estimates of components of covariance. *Biometrika*, **56**, 517–526.

Russell, T. S., and R. A. Bradley. (1958). One-way variances in the two way classification. *Biometrika*, **45**, 111–129.

Sahai, H. (1974). On negative estimates of variance components under finite population models. *S. Afr. Stat. J.*, **8**, 157–168.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.*, **2**, 110–114.

Schauder, M., and F. Schmid. (1986). How robust is one way ANOVA with respect to within group correlation? COMPSTAT: Proceedings in Computational Statistics, 7th symposium held at Rome.

Scheffe, H. (1959). *The Analysis of Variance*. John Wiley & Sons, New York.

Schultz, E. F. Jr. (1955). Rules of thumb for determining expectations of mean squares in analysis of variance. *Biometrics*, **11**, 123–135.

Searle, S. R. (1956). Matrix methods in variance and covariance components analysis. *Ann. Math. Stat.*, **27**, 737–748.

Searle, S. R. (1958). Sampling estimates of components of variance. *Ann. Math. Stat.*, **29**, 167–178.

Searle, S. R. (1961). Estimating the variability of butterfat production. *J. Agric. Sci.*, **57**, 289–294.

Searle, S. R. (1966). *Matrix Algebra for the Biological Sciences*. John Wiley & Sons, New York.

Searle, S. R. (1968). Another look at Henderson's method of estimating variance components. *Biometrics*, **24**, 749–788.

Searle, S. R. (1969). Variance component estimation: proof of equivalence of alternative method 4 to Henderson's method 3. Paper No. BU-260-M in the Mimeo Series, Biometrics Unit, Cornell University, Ithaca, NY.

Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components. *Biometrics*, **26**, 505–524.

Searle, S. R. (1971a). *Linear Models*, 1st edition. John Wiley & Sons, New York.

Searle, S. R. (1971b). Topics in variance component estimation. *Biometrics*, **27**, 1–76.

Searle, S. R. (1995). An overview of variance component estimation. *Metrika*, **42**, 215–230.

Searle, S. R. (2006). *Matrix Algebra Useful for Statistics*. John Wiley & Sons, New York.

Searle, S. R., and R. F. Fawcett. (1970). Expected mean squares in variance components models having finite populations. *Biometrics*, **26**, 243–254.

Searle, S. R., and W. H. Hausman. (1970). *Matrix Algebra for Business and Economics*. John Wiley & Sons, New York.

Searle, S. R., and C. R. Henderson. (1961). Computing procedures for estimating components of variance in the two-way classification mixed model. *Biometrics*, **17**, 607–616. Corrigenda: *Biometrics*, **23**, 852 (1967).

Searle, S. R., and J. G. Udell. (1970). The use of regression on dummy variables in management research. *Manag. Sci.*, **16**, B397–B409.

Searle, S. R., G. Casella, and C. E. Mc Culloch. (1992). *Variance Components*. John Wiley & Sons, New York.

Seelye, C. J. (1958). Conditions for a positive definite quadratic form established by induction. *Am. Math. Mon.*, **65**, 355–356.

Shah, K. R., and Y. Dodge. (1977). On the connectedness of designs. *Sankya B*, **39**(3), 284–287.

Shah, K. R., and C. G. Khatri. (1973). Connectedness in row column designs. *Commun. Stat.*, **2**(6), 571–573.

Smith, H., Jr. (1988). Stepwise regression. In: *Encyclopedia of Statistical Sciences* (9 vols. plus Supplement), vol. 8, D. L. Banks, C. B. Read, and S. Kotz (Eds.), pp. 764–768. John Wiley & Sons, Hoboken, NJ.

Snedecor, G. W., and W. G. Cochran. (1967). *Statistical Methods*, 6th edition. Iowa State University Press, Ames, IA.

Speed, F. M. (1969). A new approach to the analysis of linear models. Ph.D. Thesis, Texas A&M University, College Station, TX.

Steel, R. G. D., and J. H. Torrie. (1960). *Principles and Procedures of Statistics*. McGraw-Hill, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics and Probability*, pp. 197–206. University of California Press.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Stat. Math.*, **16**, 155–160.

Stewart, F. (1963). *Introduction to Linear Algebra*. Van Nostrand, New York.

Swallow, W. H. (1981). Variances of locally minimum variance quadratic unbiased estimators (MIVQUE) of variance components. *Technometrics*, **23**(3), 271–283.

Swallow, W. H., and J. F. Monahan. (1984). Monte Carlo comparison of ANOVA, MINQUE, REML, and ML estimators of variance components. *Technometrics*, **26**(1), 47–57.

Swallow, W. H., and S. R. Searle. (1978). Minimum variance quadratic unbiased estimation of variance components. *Technometrics*, **20**(3), 265–272.

Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Stat. Res. Mem.*, **2**, 126–149.

Theil, H., and A. S. Goldberger. (1961). On pure and mixed estimation in economics. *Int. Econ. Rev.*, **2**, 65–78.

Thompson, R. (1969). Iterative estimation of variance components for non-orthogonal data. *Biometrics*, **25**, 767–773.

Thompson, W. A., Jr. (1961). Negative estimates of variance components: an introduction. *Bull. Int. Inst. Stat.*, **34**, 1–4.

Thompson, W. A., Jr. (1962). The problem of negative estimates in variance components. *Ann. Math. Stat.*, **33**, 273–289.

Thompson, W. A., Jr. (1963). Precision in simultaneous measurement procedures, *J. Am. Stat. Assoc.*, **58**, 474–479.

Thompson, W. A., Jr., and J. R. Moore. (1963). Non-negative estimates of variance components. *Technometrics*, **5**, 441–450.

Tiao, G. C., and G. E. P. Box. (1967). Bayesian analysis of a three-component hierarchal design model. *Biometrika*, **54**, 109–125.

Tiao, G. C., and W. Y. Tan. (1965). Bayesian analysis of random-effects model in the analysis of variance. I. Posterior distribution of variance components. *Biometrika*, **52**, 37–53.

Tiao, G. C., and W. Y. Tan. (1966). Bayesian analysis of random-effects model in the analysis of variance. II. Effect of autocorrelated errors. *Biometrika*, **53**, 477–495.

Townsend, E. C. (1968). Unbiased estimators of variance components in simple unbalanced designs. Ph.D. Thesis, Cornell University, Ithaca, NY.

Townsend, E. C. (1969). Lecture notes, Statistics 363, University of West Virginia, Morgantown, WV.

Townsend, E. C., and S. R. Searle. (1971). Best quadratic unbiased estimation of variance components from unbalanced data in the 1-way classification. *Biometrics*, **27**, 643–657.

Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**(3), 232–242.

Urquhart, N. S. (1968). Computation of generalized inverses which satisfy special conditions. *SIAM Rev.*, **10**, 216–218.

Urquhart, N. S. (1969). The nature of the lack of uniqueness of generalized inverse matrices. *SIAM Rev.*, **11**, 268–271.

Urquhart, N. S., D. L. Weeks, and C. R. Henderson. (1970). Estimation associated with linear models: a revisitation. Paper BU-195, Biometrics Unit, Cornell University, Ithaca, NY.

Urquhart, N. S., D. L. Weeks, and C. R. Henderson. (1973). Estimation associated with linear models: a revisitation. *Commun. Stat.*, **1**, 303–330.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of variances is large. *Trans. Am. Math. Soc.*, **54**, 426–482.

Wang, Y. Y. (1967). A comparison of several variance component estimators. *Biometrika*, **54**, 301–305.

Weeks, D. L., and D. R. Williams. (1964). A note on the determination of connectedness in an *N*-way cross classification. *Technometrics*, **6**, 319–324. Errata: *Technometrics*, **7**, 281 (1965).

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*. **38**, 330–336.

Welch, B. L. (1956). On linear combinations of several variances. *J. Am. Stat. Assoc.*, **50**, 132–148.

Wilk, M. B., and O. Kempthorne. (1955). Fixed, mixed and random models. *J. Am. Stat. Assoc.*, **50**, 1144–1167.

Wilk, M. B., and O. Kempthorne. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Stat.*, **27**, 950–985.

Williams, E. J. (1959). *Regression Analysis*. John Wiley & Sons, New York.

Wilson, R. (1993). Review of development and application of CRSTER and MPTER models. *Atmos. Environ.*, **27**, 41–57.

Winer, B. J., D. R. Brown, and K. M. Michels. (1991). *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, New York.

Woodbury, M. A. (1950). Inverting modified matrices. Statistical Research Group, Memo, Rep. No. 42, Princeton University, Princeton, NJ.

Wright, R. L., and S. R. Wilson. (1979). On the analysis of soil variability with an example from Spain. *Geoderma*, **22**, 297–313.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire J. Exp. Agric.*, **1**, 129–142.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Stat. Assoc.*, **29**, 51–66.

Young, C. W., J. E. Legates, and B. R. Farthing. (1965). Pre-and post-natal influences on growth, prolificacy and maternal performance in mice. *Genetics*, **52**, 553–561.

Zelen, M. (1968). Discussion of Searle [1968]. *Biometrics*, **24**, 779–780.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno di Finetti*, P. Goel and A. Zellner (Eds.). North Holland, Amsterdam.

Zyskind, G., and F. B. Martin. (1969). A general Gauss-Markov theorem for linear models with arbitrary non-negative covariance structure. *SIAM J. Appl. Math.*, **17**, 1190–1202.

# AUTHOR INDEX




*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

# SUBJECT INDEX




*Linear Models*, Second Edition. Shayle R. Searle and Marvin H. J. Gruber.

# WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.