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

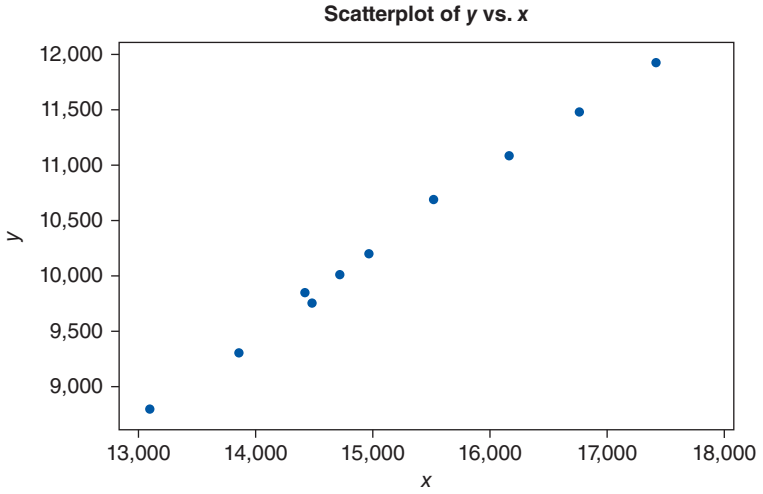
There are many practical real-world problems in many different disciplines where analysis using linear models is appropriate. We shall give several examples of such problems in this chapter as a motivation for the material in the succeeding chapters.

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

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