APPLIED LINEAR STATISTICAL MODELS

Second Edition

John Neter
William Wasserman
Michael H. Kutner

Applied Linear Statistical Models

Regression, Analysis of Variance, and Experimental Designs

John Neter
University of Georgia

William Wasserman
Syracuse University

Michael H. Kutner
Emory University

1985 Second edition

IRWIN
Homewood, Illinois 60430

Preface

Linear statistical models for regression, analysis of variance, and experimental designs are widely used today in business administration, economics, and the social, health, and biological sciences. Successful applications of these models require a sound understanding of both the underlying theory and the practical problems that are encountered in using the models in real-life situations. While *Applied Linear Statistical Models*, Second Edition, is basically an applied book, it seeks to blend theory and applications effectively, avoiding the extremes of presenting theory in isolation and of giving elements of applications without the needed understanding of the theoretical foundations.

The second edition differs from the first in a number of important respects.

1. We have added some important new topics. In the area of regression models, we have introduced discussions of detection of multicollinearity, ridge regression, and detection of influential observations. In recent years, noteworthy new developments in the detection of multicollinearity and influential observations have taken place, and a current text in linear models needs to cover these topics adequately. We have also added a chapter on nonlinear regression to introduce the reader to nonlinear statistical models.

In the area of experimental designs, we have added two chapters on nested designs, including rules for finding expected mean squares and sums of squares for a wide variety of balanced nested and crossed designs.

We have also added some specialized topics, such as P-values of test statistics, the method of least absolute deviations, analysis of variance techniques when the treatment means have unequal importance, and single degree of freedom tests.

2. We have reorganized and expanded a number of topics. In the area of

regression analysis, these include weighted regression, selection of independent variables, and normal probability plots. In analysis of variance, we have greatly expanded the discussion of the unbalanced case, devoting a full chapter to the two-factor unbalanced case including a consideration of missing cells. In the discussion of experimental designs, we have unified the treatment of missing observations in terms of the regression approach, and have split the material on randomized block designs into two chapters to facilitate understanding. At the same time we have made extensive revisions of other materials on the basis of classroom experience to improve the clarity of the presentation.

3. We have strengthened the integration of regression analysis, analysis of variance, and experimental designs through the general linear model, and have simplified the discussion by considering the same coding of indicator variables for all analysis of variance models.

4. We have placed greater emphasis on graphic methods for diagnosis, analysis, and presentation of results, and have introduced a number of computer-generated plots to demonstrate the usefulness of computer graphics.

5. The scope of the examples has been expanded to include applications from the health and biological sciences, in addition to applications from management, economics, and the social sciences. In all cases, an application can be readily understood by the general reader, regardless of background.

6. We have added three extensive real-world data sets that can be employed in a variety of ways.

7. Finally, we have substantially expanded the problem materials at the ends of the chapters and have grouped them into three categories, namely *Problems, Exercises,* and *Projects*. The *Problems* category includes basic problems and questions, the *Exercises* category includes conceptual and theoretical questions, and the *Projects* category includes problems utilizing large data sets and/or involving extensive calculations and analysis.

The first 15 chapters of the Second Edition of *Applied Linear Statistical Models* have also been published as a separate book under the title *Applied Linear Regression Models*.

A key feature of *Applied Linear Statistical Models* is its unified approach to the application of linear statistical models in regression, analysis of variance, and experimental designs. Instead of treating these areas in isolated fashion, we seek to show the interrelationships between them. Use of a common notation for regression on the one hand and analysis of variance and experimental designs on the other facilitates a unified view. The notion of a general linear statistical model, which arises naturally in the context of regression models, is carried over to analysis of variance and experimental design models to bring out their relation to regression models. This unified approach also has the advantage of simplified presentation.

We have included in this book not only the more conventional topics in regression, analysis of variance, and basic experimental designs, but also have taken up topics that are frequently slighted though important in practice. Thus, we devote a full chapter to indicator variables, covering both dependent and independent indicator variables. Another chapter takes up computer-assisted selection procedures for obtaining a "best" set of independent variables to be employed in the regression model. The use of residual analysis for examining the aptness of the model is a recurring theme throughout this book. So is the use of remedial measures that may be helpful when the model is not appropriate. In the analysis of the results of a study, we emphasize the use of estimation procedures, rather than tests, because estimation is often more meaningful in practice. Also, since practical problems seldom are concerned with a single comparison, we stress the use of multiple comparison procedures.

Theoretical ideas are presented to the degree needed for good understanding in making sound applications. Proofs are given in those instances where we feel they serve to demonstrate an important method of approach. Emphasis is placed on a thorough understanding of the models, particularly the meaning of the model parameters, since such understanding is basic to proper applications. A wide variety of case examples is presented to illustrate the use of the theoretical principles, to show the great diversity of applications of linear statistical models, and to demonstrate how analyses are carried out for different problems.

We use "Notes" and "Comments" sections in each chapter to present additional discussion and matters related to the mainstream of development. In this way, the basic ideas in a chapter are presented concisely and without distraction. Similarly, optional "Topics" chapters supplement chapters containing the main development and present a variety of additional topics that in most cases can be omitted without loss of continuity.

Applications of linear statistical models frequently require extensive computations. We take the position that a computer is available in most applied work. Further, almost every computer user has access to program packages for regression analysis and analysis of variance of different types. Hence, we explain the basic mathematical steps in fitting a linear statistical model but do not dwell on computational details. This approach permits us to avoid many complex formulas and enables us to focus on basic principles. We make extensive use in this text of computer capabilities for performing computations and illustrate a variety of computer printouts and explain how these are used for analysis.

A selection of problems is provided at the end of each chapter (excepting Chapter 1). Here the reader can reinforce his or her understanding of the methodology and use the concepts learned to analyze data. We have been careful to supply data-analysis problems that typify genuine applications. In most problems the calculations are best handled on a calculator or computer, and we urge that this avenue be used when possible.

We assume that the reader of *Applied Linear Statistical Models* has had an introductory course in statistical inference, covering the material outlined in Chapter 1. Should some gaps in the reader's background exist, he or she can read the relevant portions of an introductory text, or the instructor of the class may use supplemental materials for covering the missing segments. Chapter 1 is primarily intended as a reference chapter of basic statistical results for continuing use as the reader progresses through the book.

Calculus is not required for reading *Applied Linear Statistical Models*. In a number of instances we use calculus to demonstrate how some important results are obtained, but these demonstrations are confined to supplementary comments or notes and can be omitted without any loss of continuity. Readers who do know calculus will find these comments and notes in natural sequence so that the benefits of the mathematical developments are obtained in their immediate context. Some basic elements of matrix algebra are needed for linear models, in general, and for multiple regression, in particular. Chapter 6 introduces these elements of matrix algebra in the context of simple regression for easy learning.

Applied Linear Statistical Models is intended for use in undergraduate or graduate courses in linear statistical models and in second courses in applied statistics. The extent to which material presented in this text is used in a particular course depends upon the amount of time available and the objectives of the course. Some possible courses include:

1. A two-quarter or two-semester course in regression, analysis of variance, and basic experimental designs might be based on the following chapters:
   Regression: 2, 3, 4, 5 (Sections 5.1–5.4), 6 (Sections 6.1–6.12), 7, 8, 10 (Sections 10.1–10.3), 12.
   Analysis of variance: 16, 17, 18, 20, 21, 25.
   Experimental designs: 26, 27, 28, 31.
2. A one-quarter or one-semester course in regression analysis might be based on the following chapters:
   2, 3, 4, 5 (Sections 5.1–5.4), 6 (Sections 6.1–6.12), 7, 8, 9, 10 (Sections 10.1–10.3), 11 (selected topics), 12, 13, 14.
3. A one-quarter or one-semester course in analysis of variance might be based on the following chapters:
   16, 17, 18, 19 (selected topics), 20, 21, 22, 23 (selected topics), 24, 25.
4. A one-quarter or one-semester course in regression and analysis of variance might be based on the following chapters:
   Regression: 2, 3, 4, 5 (Sections 5.1–5.4), 6 (Sections 6.1–6.12), 7, 8, 10 (Sections 10.1–10.3).
   Analysis of variance: 16, 17, 18, 20, 21.
5. A one-quarter or one-semester course in basic experimental designs might be based on the following chapters:
   26, 27, 28, 29, 30, 31.

As time permits, the instructor could cover additional topics in the text.

This book can also be used for self-study by persons engaged in the fields of business administration, economics, and the social, health, and biological sciences who desire to obtain competence in the application of linear statistical models.

A book such as this cannot be written without substantial assistance from others. We are indebted to the many contributors who have developed the theory and practice discussed in this book. We also would like to acknowledge appreciation to our students who helped us in a variety of ways to fashion the method of

presentation contained herein. We are grateful to the many users of the First Edition of *Applied Linear Statistical Models* who provided us with comments and suggestions based on their teaching with this text. We are also indebted to Professors James E. Holstein, University of Missouri, and David L. Sherry, University of West Florida, who carefully reviewed *Applied Linear Statistical Models* to provide suggestions for this volume. Dr. Robert L. Vogel assisted us diligently in the checking of the manuscript, for which we are most appreciative. Michael J. Lynn prepared the computer-generated plots using a Zeta model 3600 plotter, and George Cotsonis and Shizuki Yamamoto assisted us in the checking of calculations and in other ways. Rebecca Baggett, Marcia Pittard, and Jennifer Scott ably handled the typing and other preparation of a difficult manuscript. We are most grateful to all of these persons for their help and assistance.

Finally, our families bore patiently the pressures caused by our commitment to complete this revision. We are appreciative of their understanding.

John Neter
William Wasserman
Michael H. Kutner

Contents

1

Some basic results in probability and statistics

This chapter contains some basic results in probability and statistics. It is intended as a reference chapter to which you may refer as you read this book. Sometimes, specific references to results in this chapter are made in the text. At other times, you may wish to refer on your own to particular results in this chapter as you feel the need.

You may prefer to scan the results on probability and statistical inference in this chapter before reading Chapter 2, or you may proceed directly to the next chapter.

1.1 SUMMATION AND PRODUCT OPERATORS

Summation operator

The summation operator Σ is defined as follows:

(1.1) $$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + \cdots + Y_n$$

Some important properties of this operator are:

(1.2a) $$\sum_{i=1}^{n} k = nk \qquad \text{where } k \text{ is a constant}$$

(1.2b) $$\sum_{i=1}^{n} (Y_i + Z_i) = \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} Z_i$$

(1.2c) $$\sum_{i=1}^{n} (a + cY_i) = na + c\sum_{i=1}^{n} Y_i \qquad \text{where } a \text{ and } c \text{ are constants}$$

The double summation operator $\Sigma\Sigma$ is defined as follows:

(1.3) $$\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} = \sum_{i=1}^{n} (Y_{i1} + \cdots + Y_{im})$$
$$= Y_{11} + \cdots + Y_{1m} + Y_{21} + \cdots + Y_{2m} + \cdots + Y_{nm}$$

An important property of the double summation operator is:

(1.4) $$\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}$$

Product operator

The product operator Π is defined as follows:

(1.5) $$\prod_{i=1}^{n} Y_i = Y_1 \cdot Y_2 \cdot Y_3 \cdots Y_n$$

1.2 PROBABILITY

Addition theorem

Let A_i and A_j be two events defined on a sample space. Then:

(1.6) $$P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$$

where $P(A_i \cup A_j)$ denotes the probability of either A_i or A_j or both occurring; $P(A_i)$ and $P(A_j)$ denote, respectively, the probability of A_i and the probability of A_j; and $P(A_i \cap A_j)$ denotes the probability of both A_i and A_j occurring.