# Applied Linear Regression Models

John Neter
William Wasserman
Michael H. Kutner

Second Edition

# Applied Linear Regression Models

SECOND EDITION

**John Neter**
*University of Georgia*

**William Wasserman**
*Syracuse University*

**Michael H. Kutner**
*Emory University*

**IRWIN**
Burr Ridge, Illinois
Boston, Massachusetts
Sydney, Australia

Sponsoring editor: Richard T. Hercher, Jr.
Project editor: Ethel Shiell
Production manager: Bette Ittersagen
Cover design: Robyn Basquin
Compositor: Interactive Composition
Typeface: 10½/12 Times Roman
Printer: R. R. Donnelley & Sons Company

**LIBRARY OF CONGRESS**
**Library of Congress Cataloging-in-Publication Data**

Neter, John.
Applied linear regression models / John Neter, William Wasserman, Michael H. Kutner, —2nd ed.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-256-07068-7
1. Regression analysis. I. Wasserman, William. II. Kutner, Michael H. III. Title.
QA278.2.N46 1989
519.5′36—dc19 88–13138
CIP

*Printed in the United States of America*

7 8 9 0 DO 5 4

# Applied Linear Regression Models

*To*
*Dorothy, Ron, David*
*Cathy, Christopher, Timothy,*
*Randall, Erin, Fiona*
*Nancy, Michelle, Allison*

# Preface

Linear regression models are widely used today in business administration, economics, and the social, health, and biological sciences. Successful application of these models requires 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 Regression 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 a new chapter on binary dependent variables since logistic regression models for binary dependent variables are becoming increasingly important in many areas of application. The new Chapter 16 introduces the reader to simple and multiple logistic regression models, to the use of the method of maximum likelihood for estimating the regression parameters, and to procedures for making inferences about the regression parameters, estimating mean responses, odds, and odds ratios, and for predicting new observations. We also take up a goodness of fit test for assessing the aptness of the logistic regression function.

In addition, Chapter 12 on model building has been largely recast and greatly expanded. We develop the model-building process in detail in this chapter to integrate the many elements of this process considered in earlier chapters. We also include in this chapter a much expanded treatment of the validation of regression models.

**2.** We have expanded the discussion of regression diagnostics by including the *DFBETAS, DFFITS,* and *PRESS* measures among the diagnostic measures considered. We have also added a discussion of partial regression plots and have strengthened still further throughout the text the emphasis on the use of regression diagnostics. We also have added a discussion of the Box-Cox transformation as a remedial measure.

**3.** We have reorganized and expanded a number of topics. The discussion of weighted least squares is now unified and taken up in conjunction with multiple regression. The discussion of standard regression models has been reorganized, and the developments of extra sums of squares and multicollinearity have been strengthened by extensive reorganization. We have expanded Chapter 13 on autocorrelation by also taking up the Hildreth-Lu procedure for estimating the autocorrelation parameter and by adding a section on prediction intervals for forecasting a new observation.

**4.** We have added a third data set for use with binary dependent variables and have included a brief discussion of response surface methodology in Chapter 9 on polynomial regression.

**5.** Throughout the text, we have made extensive revisions in the exposition on the basis of classroom experience to improve the clarity of the presentation.

We have included in this book not only the more conventional topics in regression but also topics that are frequently slighted, though important in practice. Thus, we devote two full chapters to indicator variables, covering both dependent and independent indicator variables. Another chapter takes up the model-building process, including computer-assisted selection procedures for identifying "good" subsets of independent variables for thorough analysis before a final selection is made of the regression model, and validation of the chosen regression model. The use of residual analysis and other diagnostics for examining the aptness of a regression 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 significance tests, because estimation is often more meaningful in practice. Also, since practical problems seldom are concerned with a single estimate, we stress the use of simultaneous estimation 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 regression 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 regression 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.

Applications of regression 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. Hence, we explain the basic mathematical steps in fitting a regression 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.

We assume that the reader of *Applied Linear Regression 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 Regression 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 multiple regression. Chapter 6 introduces these elements of matrix algebra in the context of simple regression for easy learning.

*Applied Linear Regression Models* is intended for use in undergraduate and graduate courses in regression analysis 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. The basic elements of regression are covered in Chapters 2, 3, 4, 5 (Sections 5.1–5.3 only), 6, 7, 8, 11, and 12. Chapters 9, 10, 13, 14, 15, and 16 can be covered as time permits and interests dictate.

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 regression 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 *Applied Linear Statistical Models* and *Applied Linear Regression Models* who have 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, for their review of *Applied Linear Statistical Models,* First Edition, to Professors Samuel Kotz, University of Maryland, Ralph P. Russo, University of Iowa, and Peter F. Thall, The George Washington University, for their review of *Applied Linear Regression Models,* and to Professors John S. Y. Chiu, University of

Washington, James A. Calvin, University of Iowa, and Michael F. Driscoll, Arizona State University, for their review of *Applied Linear Statistical Models,* Second Edition. These reviews provided many important suggestions, for which we are most grateful.

George Cotsonis and Alvin H. Rampey assisted us diligently in the checking of the manuscript, in preparing computer-generated plots, and in other ways. Almost all of the typing was done by Jane Disney, who ably handled the 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

## Chapter 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 $\Sigma$ is defined as follows:

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

Some important properties of this operator are:

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

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

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

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

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

$$= Y_{11} + \cdots + Y_{1m} + Y_{21} + \cdots + Y_{2m} + \cdots + Y_{nm}$$

An important property of the double summation operator is:

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

## Product Operator

The product operator $\Pi$ is defined as follows:

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

# 1.2 PROBABILITY

## Addition Theorem

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

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

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.

## Multiplication Theorem

Let $P(A_i | A_j)$ denote the conditional probability of $A_i$ occurring, given that $A_j$ has occurred. This conditional probability is defined as follows:

$$P(A_i | A_j) = \frac{P(A_i \cap A_j)}{P(A_j)} \qquad P(A_j) \neq 0 \tag{1.7}$$

The multiplication theorem states:

$$P(A_i \cap A_j) = P(A_i)P(A_j | A_i) \tag{1.8}$$

$$= P(A_j)P(A_i | A_j)$$

## Complementary Events

The complementary event of $A_i$ is denoted by $\bar{A}_i$. The following results for complementary events are useful:

(1.9) $$P(\bar{A}_i) = 1 - P(A_i)$$

(1.10) $$P(\overline{A_i \cup A_j}) = P(\bar{A}_i \cap \bar{A}_j)$$

## 1.3 RANDOM VARIABLES

Throughout this section, except as noted, we assume that the random variable $Y$ assumes a finite number of outcomes.

## Expected Value

Let the random variable $Y$ assume the outcomes $Y_1, \ldots, Y_k$ with probabilities given by the probability function:

(1.11) $$f(Y_s) = P(Y = Y_s) \qquad s = 1, \ldots, k$$

The expected value of $Y$, denoted by $E\{Y\}$, is defined by:

(1.12) $$E\{Y\} = \sum_{s=1}^{k} Y_s f(Y_s)$$

$E\{\ \}$ is called the *expectation operator*.

An important property of the expectation operator is:

(1.13) $$E\{a + cY\} = a + cE\{Y\} \qquad \text{where } a \text{ and } c \text{ are constants}$$

Special cases of this are:

(1.13a) $$E\{a\} = a$$

(1.13b) $$E\{cY\} = cE\{Y\}$$

(1.13c) $$E\{a + Y\} = a + E\{Y\}$$

**Note**

If the random variable $Y$ is continuous, with density function $f(Y)$, $E\{Y\}$ is defined as follows:

(1.14) $$E\{Y\} = \int_{-\infty}^{\infty} Yf(Y)\,dY$$