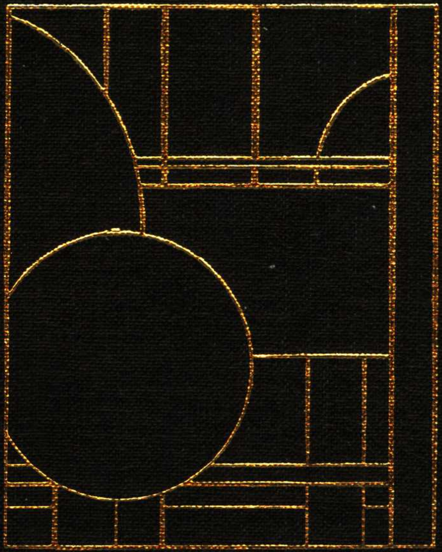
A FIRST COURSE IN LINEAR REGRESSION

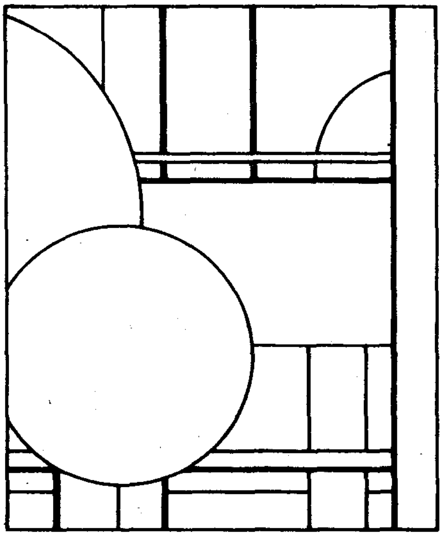
SECOND EDITION



MARY SUE YOUNGER

A FIRST COURSE IN LINEAR REGRESSION

SECOND EDITION



MARY SUE YOUNGER

THE UNIVERSITY OF TENNESSEE

DUXBURY PRESS ▪ BOSTON

PWS PUBLISHERS

Prindle, Weber & Schmidt • Duxbury Press • PWS Engineering •
Statler Office Building • 20 Park Plaza • Boston, Massachusetts 02116

PWS Publishers is a division of Wadsworth, Inc.

Minitab is the trademark of Minitab, Inc., 215 Pond Lab., University Park, PA 16802. Minitab output Copyright © 1984 by Minitab, Inc., and is used by permission.
SAS output, 82.4 and later releases, © 1983, SAS Institute, Inc. SAS is the registered trademark of SAS Institute, Inc., Cary, NC, USA.
SPSS^{X} and SPSS are trademarks of SPSS, Inc , for its proprietary computer software
BMDP is the registered trademark of BMDP Statistical Software, Inc , 1964 Westwood Blvd , Suite 202, Los Angeles, CA 90025

LIBRARY OF CONGRESS CATALOGING IN PUBLICATION DATA

Younger, Mary Sue.
A First Course in Linear Regression.
Rev. ed. of: Handbook for linear regression, c1979.
Includes index.
1. Regression analysis. 2. Regression analysis—
Data processing. I. Title.
AQ278.2Y68 1985 519.5'36 84-20592
ISBN 0-87150-865-6

ISBN 0-87150-865-6

Printed in the United States of America
85 86 87 88 89 — 10 9 8 7 6 5 4 3 2 1

Manuscript Editor: Alice Cheyer
Design and Coordinating: Susan London
Interior Illustration: Deborah Schneck
Typesetting: Monotype Composition Company
Cover Printing: New England Book Components, Inc.
Printing and Binding: Halliday Lithograph, Arcata Graphics

PREFACE

A First Course in Linear Regression introduces the techniques and applications of linear regression and provides a sound basis for more advanced study. It can also be used as a manual for those who use regression techniques without the benefit of a formal course.

It is expected that some readers may already have had an introductory survey course in regression analysis; nevertheless, this book begins at an elementary level. Building step-by-step on these elementary concepts, it offers a thorough—yet readable—explanation of regression. The text is "user-friendly," favoring verbal explanations rather than mathematical ones. Statistical theory and advanced mathematical techniques are kept to a minimum. (Theoretical assumptions and the derivations of formulas are explained intuitively in relation to commonsense considerations.) The proper application of statistical results is contrasted with cases of their incorrect use. The clear and descriptive notation conforms to that most widely used.

- The text has an early introduction to the matrix formulation of the regression problem (expanded in this revision). This serves two very important purposes:

(1) It provides a unified framework for simple, polynomial, and multiple regression and thus saves the reader from being overwhelmed by special formulas;

and

(2) It provides a basis from which to advance to the analysis of variance and covariance and to multivariate statistics.

- New material has been added on exponential functions and on dummy variables and interactions.

- *A First Course in Linear-Regression* provides step-by-step instruction in modern regression and also provides step-by-step analysis techniques using four widely available computer programs: BMDP, Minitab, SAS, and SPSS^X. Many printouts and plots are examined in detail to guide the reader in applying this software to regression analysis. Since it is expected that the reader will use only one of these four packages, each computer section stands on its own, and exercise problems may be solved using any one of the four programs. Although the matrix algebra formulation and the instructions for computer use provide many benefits to the reader, coverage of these topics can be minimized without loss of understanding if course length requires it.

- The first nine chapters discuss the simple linear regression model and introduce the matrix algebra formulation and the computer programs with regression diagnostics. Integrated with this discussion are topics presumably introduced, but seldom well-understood, in previous courses: measuring variability, constructing confidence intervals, and testing hypotheses. A single example is continued throughout the first nine chapters so that attention can be focused on new concepts rather than on new data. Also, results obtained algebraically, by matrix algebra, or from the computer, become familiar through repeated use, reinforcing the equivalence of these procedures.

- Chapter 10 serves as a bridge between single-predictor and multiple-predictor models. Fitting exponential models is a simple alteration to fitting straight-line models. Polynomial models are treated as an extension of simple linear models. Both exponential and polynomial models are reviewed. In connection with polynomial models, simultaneous, sequential, and partial tests of hypotheses are introduced.

- Chapters 11-16 discuss multiple regression. An extension of the earlier example helps illustrate concepts when more than one predictor is used. The multiple regression model is introduced using algebraic formulas, and the matrix formulation and computer programs are illustrated in detail. Simultaneous, sequential, and partial tests are developed, and the concepts of collinearity and partial correlation are explained. Dummy variables are discussed. A data set from an actual marketing study is used to illustrate various variable-selection and model-selection techniques.

- An ample number of exercises are provided at the end of each section. Many exercises introduced early are referred to later, so that the reader may utilize results already obtained.

- A large data set is included in the Appendix. Though the reader may not care to analyze the entire set, even by computer, these data may be useful in illustrating variability by drawing a random sample of cases. Also in the Appendix are "Analyses of Data Sets," which include complete numerical analyses for all exercise data sets, "Answers to Selected Exercises," and supplementary topics such as the binary number system and calculus-based derivations. *A First Course in Linear Regression* will be a valuable learning and reference source for the reader.

ACKNOWLEDGMENTS

There are many people to thank for their help in the preparation of this edition, in addition to all those listed in the preface of the first edition. Judy Lovely and Nancy Tipton did a great job of typing. Barbara Lawson attended to thousands of details, such as copying, mailing, and typing correspondence. Susan London of PWS and copy editor Alice Cheyer did an excellent job of preparing my manuscript for publication.

The University of Tennessee Computing Center has always cheerfully provided the most valuable and expert help. *A First Course in Linear Regression* has benefited especially from the advice of Paul Wright, Don Broach, and Bob Muenchen of the University of Tennessee Computing Center.

Professor John C. Rennie, Department of Forestry, Wildlife, and Fisheries, University of Tennessee, Knoxville, was kind enough to provide real data concerning the application of regression to the study of forestry. Professor David W. Cravens, formerly of the Department of Marketing at University of Tennessee, Knoxville but now at Texas Christian University, was kind enough to allow me to continue to use his data, which are so marvelously effective in illustrating variable selection procedures.

Dean C. Warren Neel of the College of Business Administration at University of Tennessee, Knoxville, and Professors Charles C. Thigpen and David L. Sylwester, former and present Chairman of the Statistics Department, respectively, have provided the facilities and resources to make this text possible. Finally, thanks to those of my colleagues and students who have been kind enough to say nice things about the first edition.

Mary Sue Younger

CONTENTS

ONE

TWO

SEVEN

INFERENCES CONCERNING THE REGRESSION COEFFICIENTS 211

EIGHT

CORRELATION 247

NINE

APPROPRIATENESS OF THE MODEL 264

TWELVE
CORRELATION IN MULTIPLE REGRESSION 425

THIRTEEN
TESTS OF SIGNIFICANCE IN MULTIPLE REGRESSION 435

FOURTEEN
DUMMY VARIABLES 458

FIFTEEN
SOME VARIABLE-SELECTION PROCEDURES 479

SIXTEEN
ANALYSIS OF RESIDUALS IN MULTIPLE REGRESSION 574

ONE

SIMPLE LINEAR REGRESSION

1.1

INTRODUCTION

In this chapter, we define regression and some of its basic terms and give some examples of situations in which regression might be the appropriate method of statistical analysis. The hypothetical data problem introduced here will be referred to throughout most of the text.

1.2

WHAT IS REGRESSION?

USES OF REGRESSION ANALYSIS

Regression

Regression is the study of relationships among variables. One purpose of regression may be to predict, or estimate, the value of one variable from known or assumed values of other variables related to it. For example, the Dean of Admissions of a university might use the College Board scores, high school class standings, personal recommendations, and scores on various tests to predict the college performance of applicants. Economists use such measures as the Index of Industrial Production and the Consumer Price Index to predict the unemployment rate or the Gross National Product. A psychologist working with laboratory rats might use number of hours of food deprivation to predict how long it will take a rat to learn its way through a maze to find food.

In order to make predictions, or estimates, we must identify the effective predictors of the variable of interest. Thus, one of the most crucial tasks in a regression study is to determine which variables are important indicators, which carry only a little information, and which predictors are redundant with other variables. Researchers can often identify many variables that may be important predictors, but they should narrow these down to the most effective predictors that can be measured at least cost.

Often a variable is difficult or very expensive to measure. For example, suppose a political scientist wanted to measure "degree of political conservatism" among a group of people. Such an intangible concept might be measured after extensive interviews on some arbitrary scale involving the subjective judgment of the interviewer. But the task would be simpler if the political scientist could find some easily measurable variables, such as income, years of education, or age, that could be used effectively to estimate degree of political conservatism.

Note that in regression analysis we use the terms *predict* and *estimate* almost interchangeably. *Predict* normally implies the future, but in regression analysis we are usually not trying to predict change over time. For example, the political scientist may find that one can use age, income, and education to estimate people's degree of political conservatism. But a different analysis would be needed to predict the degree of conservatism the same people will have as they grow older, increase or decrease in income, or go back to school. Predicting a change over time, such as next year's sales for a given company or next year's GNP, is not achieved with regression analysis. Rather, it requires *time series analysis*. While there are some similarities between time series analysis and regression analysis, the goal in regression analysis is to identify variables that carry information about another variable and not to extrapolate from present conditions to future conditions. Thus, in regression studies a **predictor** is a variable that is used to estimate some **response variable.** We expect two items that have different measurements on the predictor to have different measurements on the response also.

Predictor

Response Variable

Beyond merely identifying which variables can be used to estimate the value of another variable, regression analysis can also describe the manner in which these variables are related. One of the concerns of scientific inquiry is to formulate theories, or models, of the relationships among variables. For example, an economist might hypothesize that as the price of a certain commodity rises, the number of units sold will decline. Techniques of regression analysis can be used to see if that theory is supported or refuted by empirical evidence.

Generally, we begin with an idea—an educated guess or a hypothesis—about how several variables might be related to another variable. With the techniques of regression analysis researchers can test their hypotheses by using empirical evidence and identifying relevant predictors. Once the form of the relationship is determined, methods of regression analysis can be used to estimate the value of the variable of interest. You can probably think of several examples from your own areas of interest in which it is important to formulate models of relationships and make estimates of one variable from values of other variables.

SOME DEFINITIONS AND NOTATION

Independent Variable

Dependent Variable

When setting up a problem to analyze data, it is customary to denote the predictor, or **independent variable,** by the symbol X. Then X_1 denotes the first predictor, X_2 denotes the second, and X_k denotes the kth, or the last. The numbering of the predictors is arbitrary. However, the response variable, or **dependent variable**—the one to be predicted from the others—is usually denoted by Y. On the other hand, an economist interested in predicting the quantity of an item sold from the price per unit might call the predictor (price) P and the response (quantity) Q. Some prefer to denote the dependent variable by X_1 and the predictors by $X_2, X_3, \ldots, X_{k+1}$. Still others use X_0 for the response. For most of this text we will use$Y$ for the response variable and X_1, $X_2, \ldots, X_k$ for the predictors.

Simple Regression

Multiple Regression

We will begin with the simplest possible case: the prediction of one variable, denoted Y, from one other variable, X. A regression using only one predictor is called a **simple regression.** Though a single predictor may oversimplify reality, our results here will extend easily to more realistic situations later. Furthermore, we are often less interested in accurate predictions than in simply finding a relationship between two variables and describing the relationship. When there are two or more predictors, the analysis is called a **multiple regression.**

A SINGLE-PREDICTOR EXAMPLE PROBLEM

Suppose a new family in a certain city wants to buy a house. Aware that property values differ considerably from city to city, the newcomers seek an estimate of the prices of various houses in their new hometown. Many factors influence the price of a house: the size of the house, the size of the lot, the age and condition of the house, the part of town it is in, whether or not it is air conditioned, how badly the owner wants to sell it, local economic conditions, and so on. Let us choose just one variable, size, and see if the price can be estimated from the size. Then, in our example,

$$\begin{aligned} X &= \text{size of the house, in thousands of square feet,} \\ Y &= \text{price of the house, in thousands of dollars.} \end{aligned}$$

We certainly suspect that the price of a house depends on its size to some extent. But in order to be sure that the two variables are related we must collect some data. Some preliminary research should let us determine if there is a relationship, and if so, what it is. What kind of data should we collect? Obviously, we need to look at houses of varying sizes and record their prices. Our observations then consist of a *pair* of values for each house: