DOUGLAS C. MONTGOMERY
ELIZABETH A. PECK

INTRODUCTION TO LINEAR REGRESSION ANALYSIS

Second Edition

WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

Introduction to Linear Regression Analysis

Second Edition

DOUGLAS C. MONTGOMERY
Arizona State University

ELIZABETH A. PECK
The Coca-Cola Company

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

In recognition of the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

Library of Congress Cataloging in Publication Data:

Montgomery, Douglas C.
Introduction to linear regression analysis / Douglas C.
Montgomery, Elizabeth A. Peck. —2nd ed.
p. cm. — (Wiley series in probability and mathematical
statistics. Applied probability and statistics, ISSN 0271-6356)
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-53387-4 (alk. paper)
1. Regression analysis. I. Peck, Elizabeth A., 1953–
II. Title. III. Series.
QA278.2.M65 1991
519.5′36—dc20 91-21553
CIP

Printed in the United States of America

10 9 8 7

Preface

Regression analysis is one of the most widely used statistical techniques for analyzing multifactor data. Its broad appeal results from the conceptually simple process of using an equation to express the relationship between a set of variables. Regression analysis is also interesting theoretically because of the elegant underlying mathematics. Successful use of regression analysis requires an appreciation of both the theory and the practical problems that often arise when the technique is employed with real-world data.

This book is intended as a text for a basic course in linear regression analysis. It contains the standard topics as well as some of the newer and more unconventional ones and blends both theory and application so that the reader will obtain an understanding of the basic principles necessary to apply regression methods in a variety of practical settings. This book is an outgrowth of lecture notes for a course in regression analysis. This course is typically taken by seniors and first-year graduate students in various fields of engineering, the physical sciences, applied mathematics, and management. We have also used the material in many seminars and short courses on regression for professional audiences. It is assumed that the reader has basic knowledge of statistics such as that usually obtained from a first course, including familiarity with significance tests, confidence intervals, and the normal, t, χ^2, and F distributions. Some knowledge of matrix algebra is also necessary.

The widespread availability of computers and good software continues to contribute to the expanding use of regression. This book is oriented toward the analyst who uses computers for problem solution. In the second edition we have greatly expanded the discussion of regression diagnostics, illustrating all of the major diagnostic procedures that are available in contemporary software packages.

The book is divided into 10 chapters. Chapters 1 and 2 introduce linear regression models and provide the standard results for least squares estimation in simple linear regression. Chapter 3 discusses methods for model adequacy checking, including the basic residual plots, testing for lack of fit, detection of outliers, and other diagnostics for investigating departures from the usual regression assumptions. Remedial measures for most of these problems are also discussed, including analytical methods for selecting transformations to stabilize variance and achieve linearity. Chapter 4 is an introduction to least squares fitting in multiple regression. In addition to the standard results, several techniques are presented that we have found extremely helpful in practice, including special types of residual plots, techniques for identifying high-leverage or influential subsets of data, a procedure for obtaining a model-independent estimate of error, identification of interpolation and extrapolation points in prediction, various methods for scaling residuals, and an introduction to the multicollinearity problem. Chapter 5 discusses the special case of polynomials, including a brief introduction to piecewise polynomial fitting using splines. Chapter 6 introduces modeling and analysis considerations when using indicator variables. Variable selection and model building is presented in Chapter 7. Both stepwise-type and all possible regression selection algorithms are presented along with several summary statistics for evaluating subset regression models.

The first seven chapters form the nucleus of a modern, practically oriented course in linear regression analysis. The last three chapters present a collection of topics that sometimes fall beyond the range of a first course but that are of increasing importance in the practical use of regression. Chapter 8 focuses on the multicollinearity problem. Included are the sources of multicollinearity, its harmful effects, available diagnostics, and a survey of remedial measures. While we give an extensive discussion of biased estimation, we emphasize that it is not a cureall, and its indiscriminant use should be avoided. Chapter 9 introduces several topics, including autocorrelated errors, weighted and generalized least squares, robust regression methods, generalized linear models, and an introduction to least squares for nonlinear models. Problems of multicollinearity, autocorrelation, and nonnormality occur frequently in practice, and we believe that students in a first course should be introduced to them. We postponed any extensive discussion of these topics until this point in the text because the concepts can be more easily understood once the reader has a firm grasp of ordinary least squares methods. The material on nonlinear regression is essential for engineers and scientists, and although our book focuses on linear regression, every modern regression course should include some discussion of the nonlinear regression problem. Chapter 10 introduces model validation, which, as opposed to internal adequacy checking to measure the quality of the fit, is designed to investigate the likely success of the model in its intended operating environment.

Each chapter (except Chapter 1) contains a set of exercises. These exercises include both straightforward computational problems designed to reinforce the reader's understanding of regression methodology and mind-expanding problems dealing with more abstract concepts. In most instances the problems utilize real data or are based on real-world settings that represent typical applications of regression. A computer is helpful (in some cases necessary) in solving some of these problems, and we urge the reader to take full advantage of this resource.

Introduction to Linear Regression Analysis

Contents

CHAPTER 1

Introduction

1.1 REGRESSION AND MODEL BUILDING

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

As an example of a problem in which regression analysis may be helpful, suppose that an industrial engineering employed by a soft drink beverage bottler is analyzing the product delivery and service operations for vending machines. He suspects that the time required by a route deliveryman to load and service a machine is related to the number of cases of product delivered. The engineer visits 25 randomly chosen retail outlets having vending machines, and the in-outlet delivery time (in minutes) and the volume of product delivered (in cases) are observed for each. The 25 observations are plotted in Figure 1.1a. This graph is called a *scatter diagram*. This display clearly suggests a relationship between delivery time and delivery volume; in fact, the impression is that the data points generally, but not exactly, fall along a straight line. Figure 1.1b illustrates this straight-line relationship.

If we let y represent delivery time and x represent delivery volume, then the equation of a straight line relating these two variables is

$$y = \beta_0 + \beta_1 x \tag{1.1}$$

where β_0 is the intercept and β_1 is the slope. Now the data points do not fall exactly on a straight line, so (1.1) should be modified to account for this. Let the difference between the observed value of y and the straight line ($\beta_0 + \beta_1 x$) be an *error* ε. It is convenient to think of ε as a statistical error; that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on delivery time, measurement errors, and so forth. Thus a more plausible

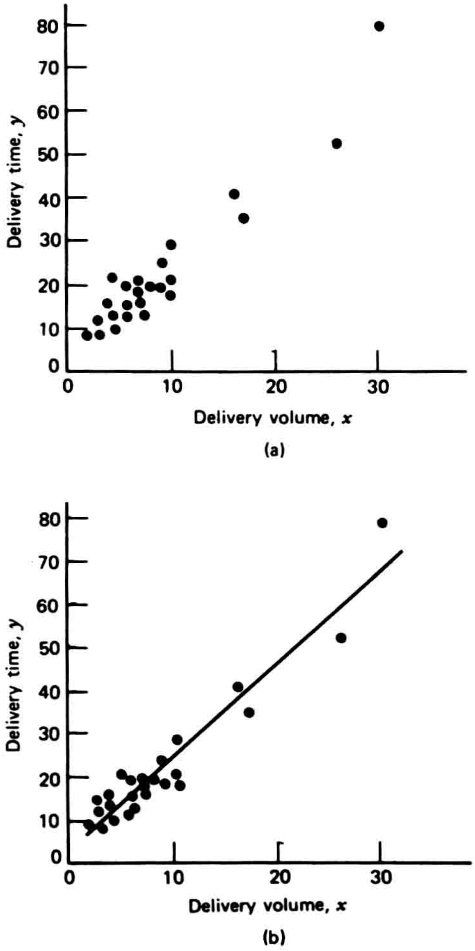

Figure 1.1 (a) Scatter diagram for delivery volume. (b) Straight-line relationship between delivery time and delivery volume.

model for the delivery time data is

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.2}$$

Equation (1.2) is called a *linear regression model*. Customarily x is called the independent variable and y is called the dependent variable. However, this often causes confusion with the concept of statistical independence, so we refer to x as the *predictor* or *regressor* variable and y as the *response* variable. Because (1.2) involves only one regressor variable, it is called a simple linear regression model. In general, the response may be related to k

regressors, $x_1, x_2, \dots, x_k$, so that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1.3}$$

This is called a multiple linear regression model because more than one regressor is involved. The adjective *linear* is employed to indicate that the model is linear in the parameters $\beta_0, \beta_1, \dots, \beta_k$, not because y is a linear function of the x's. We shall see subsequently that many models in which y is related to the x's in a nonlinear fashion can still be treated as linear regression models as long as the equation is linear in the β's.

An important objective of regression analysis is to estimate the unknown parameters in the regression model. This process is also called fitting the model to the data. We will study several parameter estimation techniques in this book. One of these techniques is the method of least squares (introduced in Chapter 2). For example, the least squares fit to the delivery time data is

$$\hat{y} = 3.3208 + 2.1762x$$

where $\hat{y}$ is the fitted or estimated value of delivery time corresponding to a delivery volume of x cases. This fitted equation is plotted in Figure 1.1b.

The next phase of a regression analysis is called *model adequacy checking*, in which the appropriateness of the model is studied and the quality of the fit ascertained. Through such analyses the usefulness of the regression model may be determined. The outcome of adequacy checking may indicate either that the model is reasonable or that the original fit must be modified. Thus regression analysis is an *iterative* procedure, in which data lead to a model and a fit of the model to the data is produced. The quality of the fit is then investigated, leading either to modification of the model or the fit or to adoption of the model. This process will be illustrated several times in subsequent chapters.

A regression model does not imply a cause–effect relationship between the variables. Even though a strong empirical relationship may exist between two or more variables, this cannot be considered evidence that the regressor variables and the response are related in a cause–effect manner. To establish causality, the relationship between the regressors and the response must have a basis outside the sample data—for example, the relationship may be suggested by theoretical considerations. Regression analysis can aid in confirming a cause–effect relationship, but it cannot be the sole basis of such a claim.

In almost all applications of regression, the regression equation is only an approximation to the true relationship between variables. For example, in Figure 1.2, we have illustrated a situation where a relatively complex relationship between y and x may be well approximated by a linear regression equation. Sometimes a more complex approximating function is necessary, as

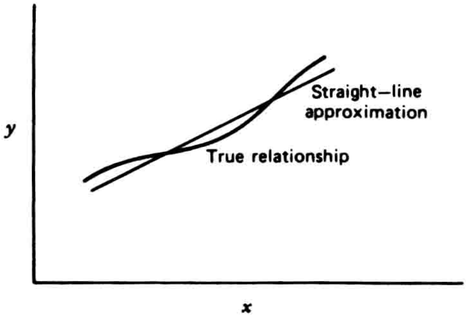

Figure 1.2 Linear regression approximation of a complex relationship.

in Figure 1.3, where a "piecewise-linear" regression function is used to approximate the true relationship between y and x.

Generally regression equations are valid only over the region of the regressor variables contained in the observed data. For example, consider Figure 1.4. Suppose that data on y and x were collected in the interval $x_1 \leqslant x \leqslant x_2$. Over this interval the linear regression equation shown in Figure 1.4 is a good approximation of the true relationship. However, suppose this equation were used to predict values of y for values of the regressor variable in the region $x_2 \leqslant x \leqslant x_3$. Clearly the linear regression model is useless over this range of x because of model error or equation error.

Finally it is important to remember that regression analysis is part of a broader data-analytic approach to problem solving. That is, the regression equation itself may not be the primary objective of the study. It is usually more important to gain insight and understanding concerning the system generating the data.

An essential aspect of regression analysis is data collection. Because the conclusions from the analysis are conditional on the data, a good data

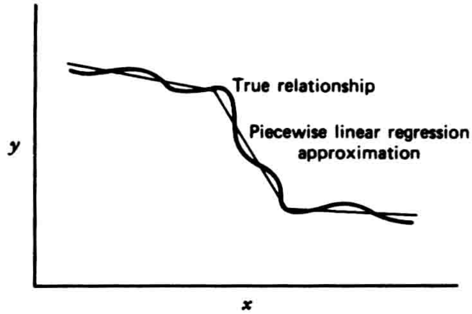

Figure 1.3 Piecewise linear approximation of a complex relationship.

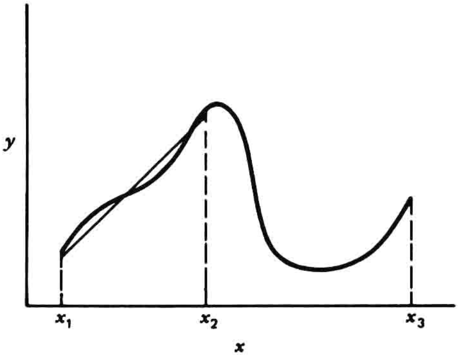

Figure 1.4 The danger of extrapolation in regression.

collection effort can have many benefits, including a simplified analysis and a more generally applicable model. The data used in a regression analysis must be representative of the system studied. Without representative data the regression model and conclusions drawn from it are likely to be in error. Care should be devoted to accurate data collection. Many of the techniques used in regression analysis can be seriously distorted by inaccurately recorded data. Preliminary data editing before the regression analysis is conducted often identifies these errors.

1.2 USES OF REGRESSION

Regression models are used for several purposes, including the following:

1. Data description.
2. Parameter estimation.
3. Prediction and estimation.
4. Control.

Engineers and scientists frequently use equations to summarize or describe a set of data. Regression analysis is helpful in developing such equations. For example, we may collect a considerable amount of delivery time and delivery volume data, and a regression model would probably be a much more convenient and useful summary of those data than a table or even a graph.

Sometimes parameter estimation problems can be solved by regression methods. For example, suppose that an electrical circuit contains an unknown resistance of R ohms. Several different known currents are passed through the circuit and the corresponding voltages measured. The scatter diagram will indicate that voltage and current are related by a straight line through the origin with slope R (because voltage E and current I are related