SANFORD WEISBERG

APPLIED LINEAR REGRESSION

WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

Applied Linear Regression

SANFORD WEISBERG

University of Minnesota
St. Paul, Minnesota

JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto

Library of Congress Cataloging in Publication Data:

Weisberg, Sanford, 1947–
Applied linear regression.

(Wiley series in probability and mathematical statistics)
Bibliography: p.
Includes index.
1. Regression analysis. I. Title.

QA278.2.W44 519.5'36 80-10378
ISBN 0-471-04419-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

APPLIED LINEAR REGRESSION

To my parents and to Carol

PREFACE

Linear regression analysis consists of a collection of techniques used to explore relationships between variables. It is interesting both theoretically because of the elegance of the underlying theory, and from an applied point of view, because of the wide variety of uses of regression that have appeared, and continue to appear every day. In this book, regression methods, used to fit models for a dependent variable as a function of one or more independent variables, are discussed for the reader who wants to learn to apply them to data. The central themes are building models, assessing fit and reliability, and drawing conclusions. If used as a textbook, it is intended as a second or third course in statistics. The only definite prerequisites are familiarity with the ideas of significance tests, p-values, confidence intervals, random variables, estimation of parameters, and also with the normal distribution, and distributions derived from it, such as Student's t, and the F, and χ^2. Of course, additional knowledge of statistical methods or linear algebra will be of value.

The book is divided into 11 chapters. Chapters 1 and 2 provide fairly standard results for least squares estimation in simple and multiple regression, respectively. The third chapter is called "Drawing Conclusions" and is about interpreting the results of the methods from the first two chapters. Also, a discussion of the effects of independent variables that are imperfectly measured is given. Chapter 4 presents additional results on least squares estimation. Chapters 5 and 6 cover methods for studying the lack of fit of a model, checking for failures of assumptions, and assessing the reliability of a fitted model. In Chapter 5, theoretical results for the necessary statistics are given, since these will be unfamiliar to many readers, while Chapter 6 covers graphical and other procedures based on these statistics, as well as possible remedies for the problems they uncover. In Chapter 7, the topics covered are relevant to problems of model building, including dummy variables, polynomial regression, and principal components. Then, Chapter 8 provides methods for selecting a model based on a subset of variables. In Chapter 9, special considerations when regression methods are to be used to make predictions are discussed. In

each of these chapters, the methods discussed are illustrated by examples using real data.

The last two chapters are shorter than the earlier ones. Chapter 10 gives guidelines for analysis of partially observed or incomplete data. Finally, in Chapter 11, alternatives to least squares estimates are discussed.

Several of the chapters have associated appendixes that have been collected at the end of the text, but are numbered to correspond to the chapters. For example: Appendix 1A.2 is the second appendix for Chapter 1. The chapters are ordered for a semester or quarter course on linear regression, and Chapters 1 to 8 make up a rigorous one-quarter course.

Homework problems are provided for each of the first nine chapters. The theoretical problems are intended only for students with the necessary statistical background. Problems that require analysis of data are intended for everyone. Some of these have been left vague in their requirements, so that they can be varied according to the interests of the students. Most of the problems use real data and can be approached in many ways.

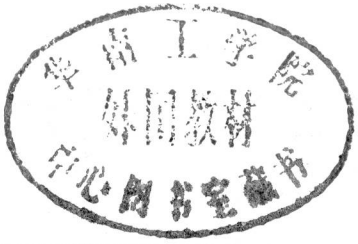
Computers. The growth of the use of regression methods can be traced directly to wider availability of computers. While this book is not intended as a manual for any specific computer program, it is oriented for the reader who expects to use computers to apply the techniques learned. High quality software for regression calculations is available, and references to the necessary sources are in the text, in the homework problems, and in the appendixes.

Acknowledgments. I am grateful to the many people who have commented on early drafts of the book, supplied examples, or through discussion have clarified my own thoughts on the topics covered. Included in this group are Christopher Bingham, Morton Brown, Cathy Campbell, Dennis Cook, Stephen Fienberg, James Frane, Seymour Geisser, John Hartigan, David Hinkley, Alan Izenman, Soren Johansen, Kenneth Koehler, David Lane, Kinley Larntz, John Rice, Donald Rubin, Wei-Chung Shih, G. W. Stewart, Douglas Tiffany, Carol Weisberg, Howard Weisberg, and an anonymous reader. Also, I wish to thank the production staff at the University of Minnesota, Naomi Miner, Sue Hangge, Therese Therrien, and especially Marianne O'Brien, whose expert assistance made completion of this work a reality.

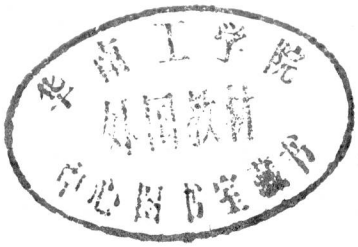
During the writing of this book, I have benefited from partial support from a grant from the U.S. National Institute of General Medical Sciences. Additional support for computations has been provided by the University Computer Center, University of Minnesota.

SANFORD WEISBERG

St. Paul, Minnesota
February 1980



CONTENTS



1

SIMPLE LINEAR REGRESSION

Regression is used to study relationships between measurable variables. Linear regression is used for a special class of relationships, namely, those that can be described by straight lines, or by generalizations of straight lines to many dimensions. These techniques are applied in almost every field of study, including social sciences, physical and biological sciences, business and technology, and the humanities. As illustrated by the examples in this book, the reasons for fitting linear regression models are as varied as are the applications, but the most common reasons are description of a relationship and prediction of future values.

Generally, regression analysis consists of many steps. To study a relationship between a number of variables, data are collected on each of a number of units or cases on these variables. In the regression models studied here, one variable takes on the special meaning of a response variable, while all of the others are viewed as predictors of the response. It is often convenient, and sometimes accurate, to view the predictor variables as having values set by the data collector, while the response is a function of those variables. A hypothesized model specifies, except for a number of unknown parameters, the behavior of the response for given values of the predictors. The model generally will also specify some of the characteristics of the failure to provide exact fit through hypothesized error terms. Then, the data are used to obtain estimates of unknown parameters. The method of estimation studied in this book is *least squares*, although there are in fact many estimation procedures. The analysis to this point is

called *aggregate analysis*, since the main purpose is to combine the data into aggregates and summarize the fit of a model to the data. The next, and equally important, phase of a regression analysis is called *case analysis*, in which the data are used to examine the suitability and usefulness of the fitted model for the relationship studied. The results of case analysis will often lead to modification of the original prescription for a fitted model, and cycling back to the aggregate analysis after modifying the data or assumptions is often necessary.

The topic of this chapter is simple regression, in which there is a single response and a single predictor. Of interest will be the specification of an appropriate model, discussion of assumptions, the least squares estimates, and testing and confidence interval procedures.

Example 1.1 Forbes' data

In the 1840s and 1850s a Scottish physicist, James D. Forbes, wanted to be able to estimate altitude above sea level from measurement of the boiling point of water. He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. In the experiments described, he studied the relationship between pressure and boiling point. His interest in this problem was motivated by the difficulty in transporting the fragile barometers of the 1840s. Measuring the boiling point would give travelers a quick way of estimating altitudes.

Forbes collected data in the Alps and in Scotland. After choosing a location, he assembled his apparatus, and measured pressure and boiling point. Pressure measurements were recorded in inches of mercury, adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature. Boiling point was measured in degrees Fahrenheit. The data for $n = 17$ locales are reproduced from an 1857 paper in Table 1.1 (Forbes, 1857).
On reviewing the data, there are several questions of potential interest. How are pressure and boiling point related? Is the relationship strong or weak? Can we predict pressure from temperature, and if so, how well?

Forbes' theory suggested that over the range of observed values the graph of boiling point versus the *logarithm* of pressure yields a straight line. Following Forbes, we take logs to the base 10, although the base of the logarithms is irrelevant for the statistical analysis. Since the logs of the pressures do not vary much, with the smallest

Table 1.1 Forbes' data, giving boiling point (°F) and barometric pressure (inches of mercury) for 17 locations in the Alps and in Scotland.

Case Number	Boiling Point (°F)	Pressure (in. Hg)	Log(Pressure)	100 × Log(Pressure)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

being 1.318 and the largest being 1.478, we shall multiply all the values of log(pressure) by 100, as given in column 5 of Table 1.1. This will avoid studying very small numbers, without changing the major features of the analysis.

A useful way to begin a regression analysis is by drawing a graph of one variable versus the other. This graph, called a *scatter plot*, can serve both to suggest a relationship, and to demonstrate possible inadequacies of it. Scatter plots can be drawn on ordinary graph paper. The x axis (or horizontal axis) is usually reserved for the variable that is to be the predictor or describer, or independent variable. In Forbes' data this is the boiling point. The y axis or the vertical axis is usually for the quantity to be modeled or predicted, often called the response or the dependent variable. In the example, the values for the y axis are $100 \times \log(\text{pressure})$. For each of the n pairs (x, y) of values in the data, a point is plotted on the graph. Although easily produced with pencil and paper, most computer programs for regression analysis will produce this plot.

The overall impression of the scatter plot for Forbes' data (Figure 1.1) is that the points generally, but not exactly, fall on a straight line

(the line drawn in Figure 1.1 will be discussed later). This suggests that the relationship between the two variables may be described (at least as a first approximation) by specifying an equation for a straight line.

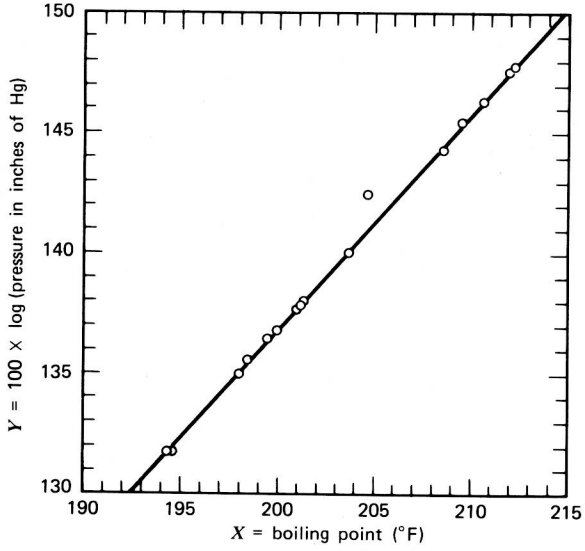


Figure 1.1 Scatter plot for Forbes' data.

As we progress through this chapter, the methods studied will be applied to these data.

1.1 Building a simple regression model

In simple regression, the relationship between two quantities, say X and Y, is studied. First, we hope that the relationship can be described by a straight line. For this to be reasonable, we may need to transform the scales of the quantities X and/or Y, as was done in Forbes' data, where pressure was transformed to log(pressure). In this chapter, observed values of the quantities X and Y are denoted by subscripted lower case letters: (x_i, y_i) are the observations on X and Y for the ith case in the study. The major features of the simple regression model are given here. A more formal approach is given in Appendix 1A.1.

Equation of a straight line. A straight line relating two quantities Y and X can be described by the equation

$$Y = \beta_0 + \beta_1 X \tag{1.1}$$

where β_0 is called the *intercept*, and corresponds to the value of Y when $X = 0$ (and is therefore the point where the line intercepts the y axis), and β_1 is called the *slope*, giving the change in Y per unit change in X (see Figure 1.2). The numbers β_0 and β_1 are called *parameters*, and, as they range over all possible values, they give all possible straight lines. In statistical applications of straight line modeling, these parameters are generally unknown, and must be estimated using the data. The difference between estimates of parameters computed from data and the actual, though unknown, values of the parameters is very important, since the data provide information about the parameters, not their actual values.

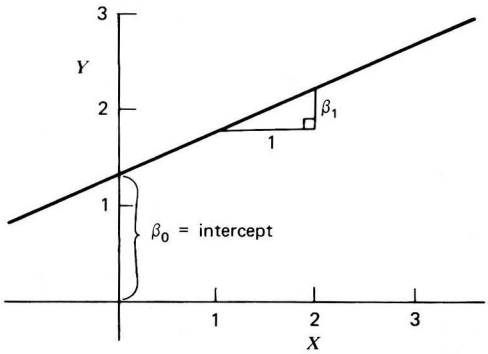


Figure 1.2 A straight line.

Errors. Real data will almost never fall exactly on a straight line. The differences between the values of the response obtained and the values given by the model (e.g., for simple regression, the observed values of Y minus $(\beta_0 + \beta_1 X)$) are called statistical *errors*. This term should not be confused with its synonym in common usage, "mistake." Statistical errors are devices that account for the failure of a model to provide exact fit. They can have both fixed and random components. A fixed component of a statistical error will arise if the proposed model, here a straight line, is not exactly correct. For example, suppose the true relationship between Y and X is given by the solid curve in Figure 1.3, and suppose that we incorrectly propose a straight line, shown as a dashed line, for this

relationship. By modeling the relationship with a straight line rather than the appropriate curve, a fixed error, sometimes called the lack of fit error, is the vertical distance between the straight line and the correct curve. For the standard linear regression theory of this chapter, we assume that the lack of fit components to the errors are negligible.

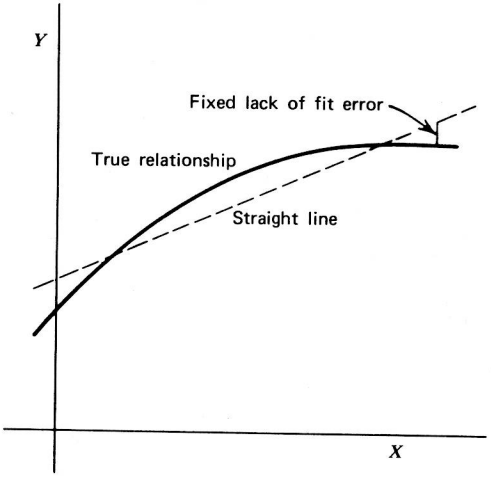


Figure 1.3 Approximating a curve by a straight line.

For the purposes of this chapter, the random component to the errors is more important. The random component can have several sources. Measurement errors (for now, consider only errors in Y, not X) are almost always present, since few quantitative variables can be measured with perfect accuracy. The effects of variables not explicitly included in the model can contribute to the errors. For example, in Forbes' experiments wind speed may have small effects on the atmospheric pressure, contributing to the variability in the observed values. Also, random errors due to natural variability occur.

Let e_i be the value of the statistical error for the ith case, $i = 1, 2, \ldots, n$. Assuming that the fixed component of the errors is negligible, the e_i have zero mean, $E(e_i) = 0$, $i = 1, 2, \ldots, n$. (See Appendix 1A.2 if the symbols $E(\ )$, var( ), and corr(,) are unfamiliar.) An additional convenient assumption is that the errors are mutually uncorrelated (written in terms of the covariance operator, as $\text{cov}(e_i, e_j) = 0$, for all $i \neq j$), and have common, though generally unknown, variance $\text{var}(e_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Heuristically, uncorrelated means that the value of one of the errors does not depend on or help determine the value of any other error. Little generality is lost if the word *independent* is substituted for *uncorrelated*. An even