LINEAR STATISTICAL MODELS AND RELATED METHODS

With Applications to Social Research

JOHN FOX

Associate Professor
Department of Sociology
York University, Toronto

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:

Fox, John, 1947–
Linear statistical models and related methods.

Includes bibliographical references and indexes.
1. Linear models (Statistics) 2. Regression
analysis. 3. Analysis of variance. 4. Social sciences--
Statistical methods. I. Title.
QA276.F6435 1984 519.5 83-23278
ISBN 0-471-09913-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*"He who loves practice without theory is like the
sailor who boards ship without a rudder and compass
and never knows where he may be cast."*

Leonardo da Vinci, 1452–1519

Preface

Linear models, their variants, and extensions are among the most useful and widely used statistical tools for social research. This book aims to provide an in-depth, modern treatment of linear models and related methods. The book should be of interest to students and researchers in the social sciences; and while the specific choice of methods and examples reflects this audience, I expect that the book will also prove useful in other disciplines that employ linear models for data analysis, and in courses on applied linear models where the subject-matter of applications is not of special concern.

My major premise in writing this text is that the teaching of social statistics should combine statistical theory, critical application, and methodology. Too often, statistical methods are presented to social scientists either as abstract formalisms or as recipes for research. I feel that there also is an unhealthy tendency to oppose statistical theory and data analysis. The quotation from Leonardo da Vinci on the relationship of theory to practice, reproduced at the front of the book, is therefore doubly apt: First, statistical methods in application must be properly related to substantive theory and research concerns. It is difficult in a statistics text to demonstrate concretely the connection between substance and statistical method, but I have attempted to select illustrations that make this point in an elementary fashion, and the connection is raised explicitly at several critical junctures. Second, I think that the sensitive use of statistical methods for data analysis requires some grounding in statistical theory. Thus, although the style of presentation in this book is nonrigorous, the statistical theory of linear models is developed alongside (or, better, underneath) its applications—few results are presented without informal derivation or intuitive justification.

Throughout the book, general approaches and principles are employed to emphasize the conceptual unity of the techniques covered. For example, the maximum-likelihood method is frequently used to derive estimators and tests, and the vector geometry of linear models is employed to clarify a variety of topics. In addition, a critical approach is adopted to the use of statistical models through explanation of (1) the assumptions underlying the models, (2) diagnostic procedures designed to assess model adequacy, and (3) means of respecifying models so that they more adequately represent the data at hand.

The book assumes that readers have been exposed to the elements of statistical inference and probability. It also assumes some knowledge of elementary calculus and matrix algebra. With the exception of elementary calculus, these areas are reviewed and extended in the appendices. Special attention is paid to topics that are frequently slighted in introductory social-statistics courses (such as probability distributions and properties of estimators), and to extensions of material presumed to be familiar (e.g., vector geometry, matrix calculus).

Chapter 1 is devoted to regression analysis, which examines the relationship of a quantitative dependent variable to one or more quantitative independent variables. Much of the statistical theory of linear models is developed in this chapter.

Chapter 2 extends linear models to include qualitative independent variables. The treatment of analysis of variance in this chapter emphasizes unbalanced (i.e., unequal-cell-frequencies) data.

Chapter 3 presents a variety of material on diagnosing and correcting linear-model problems. The problems examined include collinearity, outliers and influential data, nonlinearity, heteroscedasticity, and nonnormality. The chapter contains a discussion of data transformations and an introduction to nonlinear models.

Chapter 4 takes up structural-equation models, which are systems of linear equations representing the causal relations among sets of variables, some of which may exert mutual influence on each other. Attention is paid not only to the direct application of structural-equation models, but also to the general data-analytic principles that these models embody. The chapter ends with an introductory treatment of models that contain specific measurement-error components and that include multiple indicators of latent variables.

Chapter 5 describes logit models for qualitative dependent variables and log-linear models for contingency tables, stressing the similarity of these models to the linear models of earlier chapters. The relationship between logit and log-linear models is also developed. The chapter includes a discussion of diagnostic methods for logit models.

Most of the material in this book can be covered in a two-semester course, the first several weeks of which teach basic linear algebra and elementary differential calculus (using, e.g., Kleppner and Ramsey's *Quick Calculus*). Chapters 1 through 3 can be covered in a one-semester course that omits some specialized topics.

In learning statistics, it is important for the reader of a text to participate actively, both by working through the arguments presented in the book, and by applying methods to data. Reworking of examples is a good place to start, and I have presented illustrations in such a manner as to make re-analysis and further analysis possible.

Nearly all of the examples employ real data from the social sciences, many of them previously analyzed and published. The exercises that involve data analysis also almost all use real data drawn from a variety of areas of application. A word of warning to readers (and instructors): Many of these exercises are time consuming.[1]

Computational matters are occasionally commented on in passing, but the book generally ignores the finer points of statistical computing in favor of methods that are computationally simple. I feel that this approach facilitates learning; once basic techniques are learned, an experienced data analyst has recourse to carefully designed programs for statistical computations. Similarly, I think that it is a mistake to tie a general discussion of linear models too closely to particular programs or packages. In fact, although the marvelous proliferation of statistical software has routinized the computations for most of the methods described in this book, the workings of computer programs are not sufficiently accessible to promote learning. Consequently, I find it useful to teach APL as part of a course on linear models.

APL is an interactive programming language with powerful operators and functions, including those for common matrix operations. Similar computational facilities are available elsewhere, as in the MATRIX procedure of the SAS statistical program package and in the MINITAB package. Using APL, students are able to write their own programs for the computationally simpler methods. They may be provided with subprograms (e.g., for latent roots and vectors) with which they can construct more complex applications, and for functions such as plotting and data management.

Many of the data-analytic examples in this book are accompanied by computer-generated plots and graphs prepared on a typewriter terminal. Although these figures are less elegant than hand-drawn or computer-drawn pen-and-ink graphs, facilities for conveniently obtaining line-printer and printing-terminal plots are more widely available: I feel that it is most useful to display data as they are likely to appear to readers in their own work. The introduction of sophisticated graphics capabilities into general statistical packages (such as SAS and SPSS) and the increasing availability of flexible graphics hardware suggest that this situation will change in the future.

Because the selection of material for a text of this sort must be more than a matter of personal taste, I would like to comment briefly on some topics that are omitted from the book. First, although some attention is paid to experimentation, experimental design is not developed systematically and comprehensively: There is no discussion of nested designs, random- and mixed-effects models, variance components, and so on. These omissions reflect my judgment

[1] Starred exercises (*) are more difficult and are generally theoretical; exercises marked with a dagger (†) substantially extend the treatment in the text—readers may wish to examine these problems, even if they are not worked out in detail; and exercises marked with a pound sign (#) are intended for hand solution (i.e., with a pocket calculator rather than on a computer).

of which methods are most useful to a general social-scientific readership as well as the availability of good books on linear models for designed experiments.

Second, I have intentionally avoided a discussion of correlated errors in times-series regression. I feel that an adequate treatment of time-series data requires methods beyond the scope of this book. Time-series data are employed, however, to illustrate the problem of collinearity and in an example of a nonlinear model.

Finally, there is no systematic treatment in this book of multivariate statistical methods, although principal-components analysis is introduced in the context of collinearity, and the structural-equation models of Chapter 4 include more than one dependent variable. A serious development of multivariate statistics requires book-length treatment at a higher level of mathematical sophistication. Also, an understanding of multivariate methods is predicated on some background in univariate linear models.

JOHN FOX

Toronto, Canada
March 1984

Acknowledgments

During the 1980–1981 academic year, when the first draft of this book was written, I was partially supported by a leave fellowship from the Social Sciences and Humanities Research Council of Canada. I am grateful to the Inter-University Consortium for Political and Social Research, where I have taught for the past several summers, for providing a stimulating atmosphere in which to work.

A number of individuals have read parts of this book in draft form, and I am indebted to them for criticism and suggestions. I particularly wish to thank Ken Bollen of Dartmouth College, Gene Denzel of York University, Shirley Dowdy of the University of West Virginia, and Douglas Rivers of Harvard University for their valuable comments. Draft chapters of the book were used in courses at York University and at the ICPSR Summer Program. Students at both institutions gracefully accepted their role as guinea pigs, and their critical responses to the text—both positive and negative—helped me in the process of revision.

I wish to thank the researchers, many of them colleagues and friends, who generously made their data available to me for the examples and exercises of this book. I am also grateful to the publishers and authors who have granted me permission to reprint or adapt copyrighted material.

Finally, I wish to thank Beatrice Shube at John Wiley for her encouragement and confidence in this project.

J.F.

Contents

1

Linear Regression Analysis

This chapter develops the theory of regression analysis, a method of wide applicability and a natural point of departure for a treatment of linear models. The first section deals with simple regression, which examines the linear relationship between two numerical variables, one of which is thought to affect, influence, or predict the other. The second section takes up multiple regression, which expresses one numerical variable as a linear function of several others. The remaining sections of the chapter develop topics in linear regression analysis: random regressors (Section 1.3), model specification error (Section 1.4), and standardized regression coefficients (Section 1.5). Most of the theory described in the chapter applies generally to linear models.

When we refer to numerical variables, we mean quantitative variables measured on an interval or ratio scale. The distinction between numerical and qualitative (i.e., nominal) data is often confused with the distinction between continuous and discrete data. A numerical variable may be either continuous (such as temperature) or discrete (such as family size); a qualitative variable (sex, for example) is necessarily discrete. In data analysis, the discrete–continuous distinction becomes, for practical purposes, a distinction between variables that take on relatively few values (e.g., an individual's number of surviving grandparents) and those that take on relatively many (e.g., dollar income). Ordinal quantitative data are something of an embarrassment in linear models, and are frequently dealt with by assigning arbitrary numerical scores to their values. (This practice is discussed further in Problem 3.12, Chapter 3.)

1.1. SIMPLE REGRESSION

Simple regression analysis is a method of limited usefulness. It is, nevertheless, of critical importance to us because (1) the limitations of the method suggest directions in which it may usefully be extended, and (2) many general properties of regression analysis and linear models may be introduced in the context of simple regression.

In this section, least-squares simple regression is treated first as a method for fitting a straight line to a scatter of points that represent a set of observations for two variables. After this descriptive treatment, we introduce a statistical model for which the least-squares regression coefficients may serve as estimators. Upon developing some of the properties of the least-squares estimators, we show how these estimators may be employed for testing hypotheses about population regression coefficients. We then introduce the idea of correlation as a means of assessing the fit of the simple-regression model to data. Finally, we develop the vector geometry of simple regression analysis to provide us with a powerful conceptual tool for exploring the properties of linear models.

1.1.1. Fitting a Least-Squares Line to a Scatter of Points

Suppose that we have a set of n scores for two numerical variables X and Y. Further suppose that we regard X as a cause of Y, or at least that we wish to employ X to predict Y. Using practical terms, suppose that Y is dollar income and that X is years of education, each measured for individuals in a sample of wage earners. If X and Y were perfectly related, we could then write Y as a function of X: $Y = f(X)$. Borrowing from mathematical usage, we term X the *independent variable* and Y the *dependent variable*. Although some statisticians object to this terminology, it is quite firmly entrenched in the literature. We must be careful, however, to avoid attributing broad meaning to the terms "independent" and "dependent". For example, when we introduce several independent variables in Section 1.2, we do so without implying that these variables are *statistically* independent of one another. On occasion, we shall also refer to X as a *regressor*.

In virtually all interesting cases in the social sciences, it is unrealistic to expect that X is the sole cause of Y, or alternatively, that Y is perfectly predictable from X alone. Nevertheless, X may be one of a number of causes of Y, or Y may be partially predictable from X. There are, for instance, many causes of income beyond education: factors such as sex, race, occupation, and experience in the labor force come easily to mind. If this is the case, then we cannot expect to represent Y as an exact function of X: Two observations that are identical in their independent-variable values will in general differ in their dependent-variable values. We hope, however, to specify a function that captures the systematic relationship between the variables: $\hat{Y} = f(X)$, where $\hat{Y}$ (read "Y-hat") is called a *fitted value*. E, the difference between Y and $\hat{Y}$

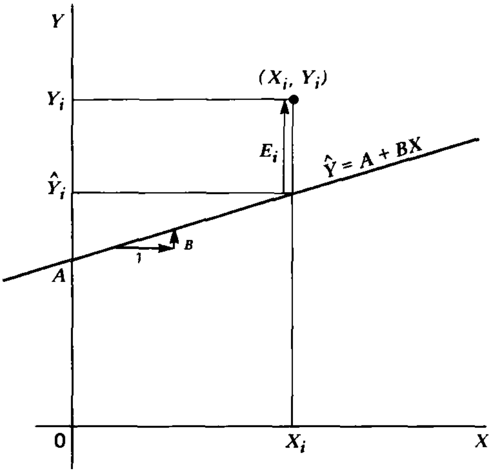

FIGURE 1.1. Linear regression of Y on X.

($E \equiv Y - \hat{Y}$), represents that portion of the dependent variable not predictable from the independent variable; E is called a *residual.* Thus, $Y = f(X) + E$. A common characteristic of the linear models developed in this text is that they decompose dependent-variable values into fitted and residual components.

Thus far, we have left the functional relation $f(X)$ between Y and X unspecified. We now specify that this relation is linear: $\hat{Y} = A + BX$; or equivalently, $Y = A + BX + E$. For the ith of our n observations, we have

$$Y_i = A + BX_i + E_i \tag{1.1}$$

Here, A and B are respectively the *Y-intercept* (or *constant*) and *slope* of the *regression line* relating Y to X, as illustrated in Figure 1.1.

Although it would be naive to suppose that the relationship between two variables is necessarily linear, straight-line regression is, for several reasons, a natural point of departure. First, it is useful to start an examination of linear models with what is reasonably regarded as the simplest interesting case.[1] Second, many relationships are in fact linear or approximately linear, or may be rendered linear by transforming the data. Finally, in the social sciences, we frequently expect one variable to increase or decrease with another without being able to specify more precisely the functional relationship between the two. It seems sensible, in these instances, to entertain linear relationships because of their simplicity. We should stress at the outset, however, that it is foolish to blindly fit relations—linear or otherwise—to data. We shall take up these issues in detail in Chapter 3.

[1] Even simpler linear models are (1) $Y = BX + E$ (regression through the origin), and (2) $Y = A + E$ (no relationship between Y and X).