Regression with Social Data

Modeling Continuous and Limited Response Variables

ALFRED DEMARIS

Bowling Green State University
Department of Sociology
Bowling Green, Ohio

A JOHN WILEY & SONS, INC., PUBLICATION

Regression with Social Data

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Regression with Social Data

Modeling Continuous and Limited Response Variables

ALFRED DEMARIS

Bowling Green State University
Department of Sociology
Bowling Green, Ohio

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

DeMaris, Alfred, 1946–
Regression with social data : modeling continuous and limited response variables / Alfred DeMaris.
p. cm. — (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 0-471-22337-9 (cloth)
1. Regression analysis. 2. Social sciences—Statistics—Methodology. 3. Statistics—Methodology. I. Title. II. Series.
HA31.3.D46 2004
519.5'36—dc22 2004041183

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Gabrielle

Contents

Preface

Here is all the invisible world, caught, defined and calculated.
In these books the Devil stands stripped of all his brute disguises.
Here are all your familiar spirits—your incubi and succubi;
your witches that go by land, by air, and by sea;
your wizards of the night and of the day.

—Arthur Miller, *The Crucible*

My students often seem to regard statistics as only slightly removed from sorcery and witchcraft. Hence I begin with the words uttered by Reverend Hale in Arthur Miller's (1954) classic play. Like Hale's books, this one also promises to demystify the arcane—in this case, regression analysis.

Regression models, in some form or another, are ubiquitous in social data analysis. Although classic linear regression assumes a continuous dependent variable, later incarnations of the technique allowed the response to take on a variety of more limited forms: binary, multinomial, truncated, censored, strictly integer, and others. Increasingly, regression texts are incorporating some limited-dependent-variable techniques—typically, binary response models—along with classic linear regression in their coverage. However, other than in econometrics texts, it is rare to find regression models for the full spectrum of continuous and limited response variables treated in one volume. This monograph aims to provide just such a treatment.

In particular, the first six chapters of the book parallel the coverage of the typical monograph on linear regression: an introduction to regression modeling (Chapter 1), simple linear regression (Chapter 2), multiple linear regression (Chapter 3), regression with categorical predictors (Chapter 4), regression with nonlinear effects (Chapter 5), and finally, a consideration of advanced topics such as generalized least squares, omitted-variable bias, influence diagnostics, collinearity diagnostics, and alternatives to ordinary least squares for heavily collinear data (Chapter 6). The second half, however, considers models for dependent variables that are limited in one way or another. Examples of such data are event counts, categorical responses, truncated responses, or censored responses. The topic coverage in the second half of the book is therefore: binary response models (Chapter 7), multinomial response models (Chapter 8), censored and truncated regression (Chapter 9), regression models for count data (Chapter 10), an introduction to survival

analysis (Chapter 11), and multistate, multiepisode, and interval-censored survival models (Chapter 12).

The book is intended both as a reference for data analysts working primarily with social data and as a graduate-level text for students in the social and behavioral sciences. As a text it is most suited to a two-course sequence in regression. As an example, I normally employ the material in Chapters 1 through 7 for a doctoral-level course on regression analysis. This course focuses primarily on linear regression but includes an introduction to binary response models. In a more advanced course on regression with limited dependent variables, I use Chapters 2 through 4 to review the multiple linear regression model, and then use Chapters 7 through 12 for the heart of the course. On the other hand, a survey of regressionlike models using the generalized linear model as the guiding framework might conceivably employ Chapters 1 through 5, and then 7 through 10. Other chapter combinations are also possible.

This book is not intended to be one's first exposure to regression. It is assumed that the reader has had a thorough introduction to probability theory, statistical inference, and applied bivariate statistics, along with an introduction to correlation and regression. Having covered the material in, say, Agresti and Finlay (1997) or Knoke et al. (2002), for example, would be good preparation for the current monograph. The basics of probability and statistical inference are nevertheless reviewed in the appendix to Chapter 1 in case the reader needs to refresh his or her understanding of these topics. It is also assumed that the reader has a solid grasp of college-level algebra. Beyond these requirements, no specialized mathematical or statistical skills are required. Some differential calculus is employed here and there in the exposition, and a smattering of matrix algebra appears—primarily in Chapter 6. Those unfamiliar with these topics will find a fairly thorough discussion of them in Appendix A. This collection of math tutorials also discusses basic algebra, summation notation, functions, and covariance algebra. These tutorials are self-contained sections that can be referred to as necessary during the course of reading through the book.

The book's emphasis tends to be on the estimation, interpretation, and evaluation of theoretically driven models in the social sciences. Due to the variety of regression models considered, coverage of specific techniques (e.g., linear regression) is necessarily more selective than found in books devoted entirely to one type of model. In particular, I have avoided discussion of exploratory model-building techniques, such as stepwise regression, along with the extensive examination of model residuals. Readers interested in these topics can find ample coverage in other works. Instead, the focus is on the substantive and statistical plausibility of models, the correct interpretation of model parameters, the global evaluation of model adequacy, and a variety of inferential procedures of interest to those working with social data. As maximum likelihood estimation is central to the models considered in Chapters 7 through 12, in the second half of the book considerable emphasis is placed on the expression for the likelihood function. This allows the reader to see how models are estimated, since once the function is written, algorithms for parameter estimation are readily available.

My writing style is the product of an attempt to marry rigor with accessibility. Rigor comes in the form of mathematical development in places where it is necessary for conveying a deeper level of understanding. Accessibility is achieved (hopefully)

by providing enough steps so that the math is clear, and by explaining the steps "in English" whenever possible. It is also hoped that the reader with more modest math skills will invest a little time and energy in the math tutorials in Appendix A. These are designed to give the reader the tools needed to at least follow the mathematical expositions in the text. As someone who developed mathematics skills rather late in life, I appreciate the trepidation with which some readers approach mathematical explication. Nonetheless, a complete understanding of this material is not possible without some math. Ideally, the returns to the reader in terms of statistical comprehension will be worth the effort.

A number of resources are available to help readers assimilate the material in the book. First, there are approximately 275 end-of-chapter exercises in Chapters 2 through 12, plus another 63 in Appendix A. The *Instructor's Manual* that accompanies the book contains complete solutions to all the exercises. Additionally, 10 datasets are available so that readers can practice the techniques taught herein using their favorite regression software. The datasets are incorporated into several of the end-of-chapter exercises. The datasets can be downloaded through the Wiley Web site, as discussed in Chapter 1 (see the section "Datasets Used in This Volume" in Chapter 1 for further information).

Acknowledgments

Many people have contributed in one way or another to the production of this work. First, I would like to thank the following statisticians for reading and commenting on preliminary book chapters: Alan Agresti, Kenneth A. Bollen, Nancy Boudreau, William H. Greene, David W. Hosmer, and James A. Sullivan. Most of these professors are people with whom I have had little or no prior connection, but who were simply gracious and collegial enough to take the time to help me produce a better product. If there are flaws in this work, it is undoubtedly due to my failure to take their sage advice. I also wish to thank Bowling Green State University for providing the resources—in particular, time and computing support—that have made it possible to complete this project in a timely fashion. Thanks are also due to my colleagues in the Department of Sociology at Bowling Green State University for being supportive of this project and for politely excusing (or at least not complaining about) my absence at department colloquia and other functions while at work on this project. Additionally, I wish to express my appreciation to Steve Quigley and the production staff at John Wiley & Sons for their professionalism as well as their encouragement of this monograph. Finally, I wish to give sincere thanks to my lovely wife, Gabrielle, for her unfailing love and support throughout the writing of this work.

CHAPTER 1

Introduction to Regression Modeling

The last several decades in the social sciences have been characterized by the increasing use of mathematical models of social behavior. The ready availability of quantitative data on social phenomena, generated by large-scale social surveys, is certainly a contributing factor in this development. Although models for social data vary widely in complexity and sophistication, most can be considered to be variants of the technique known as *linear regression*. Classic linear regression, however, was predicated on the notion that the outcome variable being modeled was continuous in nature. Many outcomes of interest, on the other hand, are limited in their measurement in some way or another. In this monograph, I define a *limited response variable* to be any outcome that is not continuous—or approximately continuous—throughout its logical range. Such measures include a continuous response that is truncated or censored, one that is categorical, and one that represents a count of some phenomenon. Also included under this definition are measures of survival time in a given state, as this type of response is also typically characterized by restrictions imposed by censoring and/or truncation. Linear regression has been extended over the years to the modeling of limited dependent variables, via the *generalized linear model*, discussed below. The purpose of this book, therefore, is to present an integrated treatment of regression modeling that weaves seamlessly through the various metrics that the response variable can take. By collecting a variety of seemingly disparate techniques under the regression umbrella, this book will hopefully render these methods easier to assimilate.

CHAPTER OVERVIEW

In this chapter I introduce the concept of a statistical model: in particular, a linear regression model. It turns out that linear regression models are special cases of what

Regression with Social Data: Modeling Continuous and Limited Response Variables,
By Alfred DeMaris
ISBN 0-471-22337-9

is referred to as the *generalized linear model* (Gill, 2001; McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972), which subsumes all the models discussed in this book. The important components of such a model are therefore sketched out in this chapter to foreshadow what is to follow in subsequent chapters. I then outline three major components of model evaluation, which are considered throughout the book for assessing model adequacy. Next, I consider the role of regression models in causal inference. Whether or not acknowledged explicitly, regression modeling in the social and behavioral sciences is frequently designed to illustrate causal dynamics. I therefore devote some space to a discussion of recent developments in, and controversies pertaining to, the use of regression models for causal inference. The chapter concludes with a description of the data sets used for this volume, some of which the reader may download to practice the techniques taught herein. Finally, the chapter appendix contains a review of important statistical principles relied on throughout the volume.

MATHEMATICAL AND STATISTICAL MODELS

In the social and behavioral sciences, a model is often a *set of one or more equations describing the processes that generated the observations on one or more response variables*. I use the term *generated* here in a causal sense, since that is what is typically implied in researchers' models, as well as the language used to describe them. (I shall have more to say about causal language shortly.) When coupled with a set of assumptions about the manner in which observations were *sampled* from a larger *population*, it becomes a *statistical model*. Like many "models" of real-world phenomena, such models are not to be taken too literally. As others have observed, "All models are wrong. Some are useful" [attributed to George Box in Gill (2001, p. 3)]. Nonetheless, to the extent that a model provides a broad outline of the dynamics underlying behavioral phenomena, it can be useful for advancing knowledge.

Linear Regression Models

A linear regression model is an *equation* in which a random response, or outcome, variable Y, is posited to be a *linear function* of a set of input, or explanatory variables, denoted X_1, X_2, ... (These labels are, of course, purely arbitrary. The outcome could just as well be denoted W, U, or η, and the explanatory variables—also called *regressors*—could be labeled V, Z, or ξ.) To give this discussion substantive flesh from the start, suppose that the "population" of interest is the population of all persons over 18 years of age in the United States in 1998. Suppose further that Y is a continuous measure of *attitude toward abortion*, with a higher score indicating a more liberal, or unrestrictive, attitude. And let's say that X_1 is *marital status*, where "married" is coded 1, and "any other status" is coded 0. (Called *dummy variables* these types of variables are explored in detail in Chapter 4.) Additionally, say that X_2 is *education*, coded from 0 for "no formal schooling" to 20 for "four or more years

of graduate study." A regression model for *attitude toward abortion* for the *i*th observation sampled from the population based on these two regressors takes the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \tag{1.1}$$

This is a linear equation, in the sense that Y is defined to be a weighted sum of constants times explanatory variables (see Sections I and II of Appendix A for definitions of functions, linear functions, and weighted sums). But—you might object—there's no variable multiplied by β_0 and no constant multiplying ε_i. Well, both are actually present. The "variable" corresponding to β_0 is X_0, which equals 1 for all cases. This factor is, therefore, easily omitted from the equation. The constant multiplier of ε_i is simply assumed to be 1. Hence, this multiplier can also be omitted. The β's—β_0, β_1, β_2—are the *parameters* of the equation: They are assumed to take on constant values for each person in the population. The last term, ε, is an equation *disturbance*, or error term. It is a random variable that represents all factors affecting Y other than X_1 and X_2. Both the parameters and ε are unobserved in any given sample. That is, even though we can observe the values of Y and the X's for any sample of n cases from the population, we cannot observe either the parameters or the error term. These factors, however, can be estimated with the sample data. In fact, the major purpose of regression modeling is to estimate the β's and to use these to describe the relationship between Y and the X's in the population, as well as to make predictions about the value of Y for cases with particular combinations of values of the X's.

Model (1.1) is for individual observations. The model for the *expected value*, or mean, or arithmetic average, of Y in the population, conditional on the X's, is instead simply

$$E(Y_i | X_{i1}, X_{i2}) = \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}. \tag{1.2}$$

The β's quantify the manner in which the mean of Y is related to the explanatory variables in the model. In particular, β_1 indicates the expected, or average, difference in Y in the population for those who are 1 unit apart in marital status—that is, for marrieds versus everybody else in our substantive example. β_2 indicates the expected difference in Y in the population for those who have a year's difference in formal schooling. So, for example, in the prediction of one's attitude toward abortion, if β_1 is -1.5 and β_2 is 2.3, these would be interpreted as follows: Married persons' attitude toward abortion, on average, is 1.5 units lower than others', holding education constant. (The precise meaning of "holding other variables constant" will be taken up in subsequent chapters.) Those with a year's more formal schooling, on average, are 2.3 units higher on attitude toward abortion than others, holding marital status constant. Furthermore, if β_0 is 7.5, a married person with a college degree is estimated to have mean abortion attitude equal to $7.5 - 1.5(1) + 2.3(16) = 42.8$.

This "model" of attitude toward abortion is certainly an oversimplification of the set of factors associated with such attitudes. But it is parsimonious, and its adequacy in accounting for variation in attitude toward abortion can be evaluated (more about this

later). To estimate the β's with sample data employing the most common technique—ordinary least squares (OLS)—we make some additional assumptions about the equation errors. First, we assume that they are uncorrelated with one another. That is, there is no tendency for a large error for the first observation, say, to presage a larger or smaller error for the second observation than would occur by chance. If sampling is random and the data are cross-sectional rather than longitudinal, this assumption is usually pretty safe. Second, we assume that they have a mean of zero at each *covariate pattern*, or combination of predictor values. As an example, being married and having 16 years of education is one covariate pattern; being other-than-married with 12 years of education is another covariate pattern; and so on. Hence, this assumption is that the mean of the errors at any covariate pattern is zero. Finally, we assume that the variance of the error terms is the same at each covariate pattern. Given a random sample of n persons from the population, along with their measures on Y, X_1, and X_2, we can proceed with an estimation of this equation and employ it to further our understanding of abortion attitudes.

Generalized Linear Model

A linear regression model is a special case of the *generalized linear model* (GLM). A generalized linear model is a *linear model for a transformed mean of a response variable whose probability distribution is a member of the exponential family* (Agresti, 2002). What does this mean? Well, for starters, let's apply this definition to the regression model delineated in equation (1.2) and corresponding assumptions above. The quantity μ_i in equation (1.2) is referred to as the *conditional mean* of the response variable. It is the mean of the Y_i conditional on a particular covariate pattern. (The ε_i are, moreover, more properly called the *conditional errors*—the errors, at each covariate pattern, in predicting the individual Y_i using the conditional mean.) The model is therefore a model for the *mean* of the response variable. It is also for the *transformed mean* of Y, although the transformation employed here is the *identity* transformation, which is "transparent" to us. That is, if $g(\mu_i)$ indicates a transformation of the mean using the function $g(\cdot)$, then $g(\mu_i)$ in the classic regression model is just μ_i. Also, in the classic regression model, it is assumed that the errors are *normally* distributed. (This assumption is not essential if n is large, however.) Because Y is a linear combination of the regressors plus the error term, and assuming that the regressor values are fixed, or held constant, over repeated sampling, Y is also normally distributed. The normal distribution is a member of the *exponential family* of probability distributions.

Essentially, there are three components that specify a generalized linear model. First, the *random* component identifies the response variable, Y, its mean, μ, and its probability distribution. Second, the *systematic* component specifies a set of explanatory variables used in a linear function to predict the transformed mean of the response variable. The systematic component, referred to as the *linear predictor* (Agresti, 2002), has the form $\sum_{k=0}^{K} \beta_k X_{ik}$ for the ith case, where the X's are the explanatory variables and the β's are the parameters representing the variables' "effects" on the mean of the response. In the example of attitude toward abortion, $\sum_{k=0}^{K} \beta_k X_{ik}$ is just $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$. Third,

the *link function*, $g(\mu)$, specifies the transformation function for the mean of Y, which the model equates to the systematic component.

The linear regression model is especially simple because the response variable is continuous—at least theoretically—and the link function is the *identity link*. That is, $g(\mu) = \mu$, and hence the regression model is $\mu_i = E(Y_i) = \sum_{k=0}^{K} \beta_k X_{ik}$, as we saw in equation (1.2). An important characteristic about this equation is that the left- and right-hand sides are equally unrestricted. That is, if Y is continuous, its theoretical range is from minus to plus infinity, which implies a similar range for μ. The right-hand side is also free to take on any values in that range, since there are no restrictions on either the parameters or the values of the predictors. However, later in this book we consider other regressionlike models in which the response variable is either binary, nonnegative discrete, or otherwise limited in its range. The link function is therefore designed to ensure that the response is converted into an unrestricted form, to match the unrestricted nature of the linear predictor. Let's consider how the GLM framework extends to those situations.

First, we need to describe the exponential family of density functions. (Readers unfamiliar with the concept of a density function may want to review that material in the chapter appendix.) A density is a member of the exponential family if it can be written in the form

$$f(y \mid \mu) = a(\mu)\, b(y) e^{y g(\mu)}, \tag{1.3}$$

where, as before, μ is the mean of Y, $a(\mu)$ is a function involving only μ, and perhaps constants, and $b(y)$ is a function involving only Y, and perhaps constants (Agresti, 2002). Once the density is written in this form, the link function that equates the mean of Y to the linear combination of explanatory variables is $g(\mu)$. As an example, suppose that the response variable, Y, is binary, taking on values 1 if a person has had sexual intercourse any time in the preceding month, and 0 otherwise. Suppose further that we are interested in modeling having had sexual intercourse in the preceding month as a function of several predictors, such as *marital status*, *education*, *age*, *religiosity*, and so on. Such a response variable is said to have the *Bernoulli distribution* with parameter π, and its density function (see the chapter appendix) is

$$f(y \mid \pi) = \pi^y (1 - \pi)^{1-y}.$$

For binary Y, $E(Y) = \pi$, so π is the mean of the response in this case. Now, since

$$\begin{aligned}
\pi^y (1 - \pi)^{1-y} &= \pi^y (1 - \pi)(1 - \pi)^{-y} \\
&= \pi^y \frac{1 - \pi}{(1 - \pi)^y} \\
&= (1 - \pi)\left(\frac{\pi}{1 - \pi}\right)^y \\
&= (1 - \pi) e^{y \ln[\pi/(1-\pi)]}
\end{aligned}$$

we see that the Bernoulli density is a member of the exponential family, with $a(\mu) = (1 - \pi)$, $b(y) = 1$, and $g(\mu) = \ln[\pi/(1 - \pi)]$. Thus, $\ln[\pi/(1 - \pi)]$ is the link function for this model, and the model for the transformed mean becomes

$$\ln \frac{\pi_i}{1-\pi_i} = \sum_{k=0}^{K} \beta_k X_{ik}.$$

This type of model is called a *logistic regression* model. Notice that since π ranges from 0 to 1, $\pi/(1 - \pi)$ ranges from 0 to infinity, and therefore $\ln[\pi_i/(1 - \pi_i)]$ ranges from minus to plus infinity. The left-hand side of this model is thus an unrestricted response, just as in the case of linear regression.

As a second example, suppose that the response on *sexual frequency* really is recorded in terms of the number of separate acts of sexual intercourse that the person has engaged in during the preceding month. This type of outcome is referred to as a *count variable*, since it represents a count of events. It is a discrete variable whose distribution is likely to be very right-skewed. We may want to utilize this information to inform the regression. One appropriate density for this type of variable is the *Poisson density*. Hence, if Y takes on values 0, 1, 2, . . . and $\mu > 0$, the Poisson density is

$$f(y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}.$$

To see that this is a member of the exponential family, we rewrite this density as

$$\frac{e^{-\mu}\mu^{y}}{y!} = e^{-\mu}\frac{1}{y!}e^{y \ln \mu},$$

where $a(\mu) = e^{-\mu}$, $b(y) = 1/y!$, and $g(\mu) = \ln \mu$. Therefore, $\ln \mu$ is the link function, and the model for the transformed mean becomes

$$\ln \mu = \sum_{k=0}^{K} \beta_k X_{ik}.$$

This model is referred to as a *Poisson regression model*. Here, in that μ ranges from 0 to infinity, $\ln \mu$ ranges from minus to plus infinity. Once again, the left-hand side of the model is an unrestricted response.

The advantage to the GLM approach is that the link function connects the linear predictor, $\sum_{k=0}^{K} \beta_k X_{ik}$, to the mean of the response variable rather than to the response variable itself, so that the outcome can now take on a variety of nonnormal forms. As Gill (2001, p. 31) states: "The link function connects the stochastic [i.e., random] component which describes some response variable from a wide variety of forms to all of the standard normal theory supporting the systematic component through the mean function, $g(\mu)$ . . . ." Once we assume a particular density function for Y, we can then employ maximum likelihood estimation (see the chapter appendix for an explanation of the maximum likelihood technique) to estimate the parameters of the model. For the classic linear regression model with

normally distributed errors (and thus a normally distributed response), maximum likelihood (ML) and ordinary least squares (OLS) estimation are equivalent (OLS estimation is covered in Chapter 2).

Model Evaluation

Models in the social sciences are useful only to the extent that they effectively encapsulate real-world processes. In this section we therefore consider ways of evaluating model adequacy. The assessment of a model encompasses three major evaluative dimensions. The first dimension is *empirical consistency*, or as many call it, *goodness of fit. A model is empirically consistent if the response variable behaves the way the model says that it should*. In other words, a model is empirically consistent to the extent that the response variable behaves in accordance with model assumptions and follows the pattern dictated by the model's structure. Moreover, if the model's predictions for Y match the actual Y values quite closely, the model is empirically consistent. The second dimension is *discriminatory power*, which is *the extent to which the structural part of the model is able to separate, or discriminate, different cases' scores on the response from one another.* Since separation, or dispersion, constitutes variability in the response, discriminatory power is typically assessed by examining how much of the variability in the response is due to the structural part of the model. The third dimension is *authenticity*, also called *model-reality consistency* by Bollen (1989). *A model is authentic to the extent that it mirrors the true processes that generated the response*.

To illustrate the differences in these dimensions, I draw on a particular variant of regression modeling called a *path model*, essentially a model for a causal system in which one or more response variables is a function of a set of predictors. A path model is an example of what is referred to as a *covariance structure model* or *structural equation model* [see DeMaris (2002a) or Long (1983) for an introduction to such models]. In this type of model, the goal is to account for the correlations (or covariances) among the variables in the system, using the structural coefficients of the model. For example, suppose that we have three continuous, standardized variables measured for a random sample of married adult respondents: Z_1 is the the degree of physical aggression in the respondent's marriage in the past year, Z_2 is the frequency of verbal disagreements in the respondent's marriage in the past year, and Z_3 is the frequency of verbal disagreements in the respondent's parents' marriage when the respondent was a teenager. The sample correlations among these variables are $\text{corr}(Z_1,Z_2) = .45$, $\text{corr}(Z_1,Z_3) = .6125$, and $\text{corr}(Z_2,Z_3) = .2756$. In path analysis, these correlations are the observations that are to be accounted for by the model.

A path model can be specified using either a diagram or a series of equations. Using the latter approach, suppose that a researcher arrives at the following OLS sample estimates for a simple path model for Z_1, Z_2, and Z_3:

$$\begin{aligned} Z_2 &= .45(Z_1) + e_2, \\ Z_3 &= .5(Z_1) + .25(Z_2) + e_3. \end{aligned} \tag{1.4}$$

The model suggests that the frequency of verbal disagreements in the respondent's marriage in the past year is a function of the degree of physical aggression in the respondent's marriage in the past year, plus a random error term (e_2). It also maintains that the frequency of verbal disagreements in the respondent's parents' marriage when the respondent was a teenager is a function of the degree of physical aggression in the respondent's marriage in the past year and the frequency of verbal disagreements in the respondent's marriage in the past year, plus a random error term (e_3). (Okay, this doesn't make much substantive sense, but *that* will be the point, as the reader can see below.) It can (and, in fact, will) be shown that the sample correlations among Z_1, Z_2, and Z_3 are functions of the model's estimated parameters. The total number of "observations" in path analysis consists of the number of nonredundant correlations among the variables in the system. In the present example, that number is three. There are also three parameters in the system: the three coefficients. Whenever the number of correlations is the same as the number of parameters in the system of equations, the model is *saturated*, or *just-identified*. In this case, the structural parameters will reproduce perfectly the correlations among the variables. When there are fewer parameters than correlations to explain, the model is *overidentified*. In that case, the model is a more parsimonious description of the correlations. The model will no longer perfectly reproduce the correlations. But we can assess how *closely* the model's parameters will reproduce the correlations in order to gauge its performance in "fitting" the data.

Let's see how the correlations can be shown to be functions of the structural parameters of the model. (Those unfamiliar with covariance algebra may want to read Section III of Appendix A before continuing.) First, note that since the variables are standardized, their covariances are also their correlations. Thus, $\text{corr}(Z_1,Z_2) = \text{cov}(Z_1,Z_2) = \text{cov}(Z_1, .45Z_1 + e_2) = .45\,\text{Cov}(Z_1,Z_1) + \text{cov}(Z_1,e_2) = .45$ (since the covariance of a variable with itself is its variance, which for standardized variables equals 1, and the covariance between OLS residuals and regressors in the same equation is zero). Moreover, $\text{corr}(Z_1,Z_3) = \text{cov}(Z_1, .5Z_1 + .25Z_2 + e_3) = .5\text{v}(Z_1) + .25\,\text{cov}(Z_1,Z_2) = .6125$; and $\text{corr}(Z_2,Z_3) = \text{cov}(.45Z_1 + e_2, .5Z_1 + .25Z_2 + e_3) = .45(.5)\,\text{v}(Z_1) + .45(.25)\,\text{cov}(Z_1,Z_2) = .2756$. (Note that OLS residuals in different equations are uncorrelated with each other.) We see that the correlations are reproduced exactly from the model parameters, because the model is saturated.

The structural coefficients also allow us to determine how much the model accounts for variation in the response variables. The part of the variance of a response variable that is accounted for by the model can be determined by considering the overall variance of each response. Recalling that the variance of a standardized variable is 1, the variance in Z_2 can be decomposed into the proportion due to the structural part of the model and the proportion due to error. Thus, we have $1 = \text{v}(Z_2) = \text{cov}(Z_2,Z_2) = \text{cov}(.45Z_1 + e_2, .45Z_1 + e_2) = .45^2\ \text{v}(Z_1) + \text{v}(e_2) = .2025 + \text{v}(e_2)$. That is, 20.25% of the variation in Z_2 is due to the structural (as opposed to the random) part of the model. Similarly, $1 = \text{v}(Z_3) = \text{cov}(.5Z_1 + .25Z_2 + e_3, .5Z_1 + .25Z_2 + e_3) = (.5)(.5)\ \text{v}(Z_1) + (.5)(.25)\ \text{cov}(Z_1,Z_2) + (.5)(.25)\ \text{cov}(Z_1,Z_2) + (.25)(.25)\ \text{v}(Z_2) + \text{v}(e_3) = .5^2 + (2)(.5)(.25)(.45) + .25^2 + \text{v}(e_3) = .425 + \text{v}(e_3)$. Here we see that 42.5% of the variation in Z_3 is due to the model.

At this point, let's consider the three aspects of model evaluation. First, notice that the model is *perfectly* empirically consistent, since the data—the correlations—"behave" exactly the way the model says they should; they are predicted perfectly by the model. Discriminatory power, on the other hand, is only moderate; at most, 42.5% of the variation in any response variable is accounted for by the model. Another way of saying this is that we experience, at most, only a 42.5% improvement in the discrimination of scores on the response variable when using—as opposed to ignoring—the model, in predicting the responses. Finally, however, the model is completely *inauthentic*, in a causal sense. To begin, the frequency of verbal disagreements in the respondent's parents' marriage when respondents were teenagers cannot possibly be caused by the subsequent tenor of respondents' marriages. Additionally, physical aggression tends to be preceded by verbal conflict rather than the converse. It is therefore unreasonable to suggest that it is physical aggression that leads to verbal conflict. If anything, the occurrence of physical aggression should suppress the frequency of subsequent verbal altercations, since partners would be fearful of a reoccurrence of violence. From the foregoing it should be clear that empirical consistency, discriminatory power, and authenticity are three separate although related criteria by which models can be evaluated.

REGRESSION MODELS AND CAUSAL INFERENCE

Regression modeling of nonexperimental data for the purpose of making causal inferences is ubiquitous in the social sciences. Sample regression coefficients are typically thought of as estimates of the causal impacts of explanatory variables on the outcome. Even though researchers may not acknowledge this explicitly, their use of such language as *impact* or *effect* to describe a coefficient value often suggests a causal interpretation. This practice is fraught with controversy [see, e.g., McKim and Turner (1997) as well as the November 1998 and August 2001 issues of *Sociological Methods & Research* for recent debates on this topic in sociology]. In this section of the chapter I explore the controversy and provide some recommendations.

What Is a Cause?

Philosophers and others have debated the definition of *cause* for centuries without ever coming to complete agreement on it. However, current common use of the term implies that the application of a cause to some element changes its state or trajectory, compared to what that would be without application of the cause. Beyond this basic idea, however, there appear to be two primary "models" of causality in operation among social scientists. The *regression* or *structural equation modeling* perspective is that *a variable X is a cause of Y if, all else equal, a change in X is followed by a change in Y* (Bollen, 1989). The implicit assumption is that a cause is synonymous with an *intervention*, which, when applied, changes the nature of the outcome, on average. With nonexperimental data, the intervention has been executed by nature. Nonetheless, the implication is that if X is truly a cause of Y, changing its

value should change Y for the cases involved, compared to what its value would be were X left unchanged. Should this reasoning be applied to equation (1.1), β_2 would be described as individuals' average change in attitude toward abortion were we to increase their schooling by one year.

A somewhat different perspective is encompassed by what is referred to as the *potential response model* of causality (Pearl, 1998), attributed to Rubin (1974), and therefore also referred to as the *Rubin model*. This viewpoint entails a counterfactual, or contrary-to-fact, requirement for causality: *X is a cause of Y if the value of Y is different in the presence of X from what it would have been in the absence of X* (*or under a different value for X*). Although this sounds quite similar to the notion of intervention articulated above, there are some subtle differences. First, let's consider the potential response model more formally. Suppose that X represents a treatment with two values: t for the treatment itself and c for the absence of treatment. Define Y_t as the score on a response, Y, for the ith case if the case had been exposed to t, and Y_c as the response for the same case if that case had *instead* been exposed to c. Then the *true causal effect* of X on Y for the ith case is $Y_t - Y_c$. Notice that this definition of cause is counterfactual, since the ith case can be "freshly" exposed to either t or c but not to both. Repeated application of c followed by t is not considered equivalent. Similarly, the *average causal effect* for some population of cases is the average of all true causal effects for all cases. That is, the average causal effect is $E(Y_t - Y_c)$ over the population of cases. Neither the true causal effect nor the average causal effect can ever be observed, in practice. Notice the difference between this model and the intervention approach to causality discussed above. An intervention is an observable operation. What's more, it is indifferent to the case's prior history: We can *change* the case's value from c to t and observe what happens, on average, to Y. The potential response model, in contrast, defines causality in a way that is impossible to observe, since the values Y_t and Y_c presume that the case's history has been magically "erased" in each case before a particular level of X is applied.

Nonetheless, according to the potential response model, the average causal effect can be estimated in an unbiased fashion if there is random assignment to the cause. Unfortunately, this pretty much rules out making causal inferences from nonexperimental data. However, others acknowledge the possibility of making the assumption of "conditional random assignment" to the cause in observational data, provided that this assumption is theoretically tenable (Sobel, 1998). Still, hard-core adherents to the potential response framework would deny the causal status of most of the interesting variables in the social sciences because they are not capable of being assigned randomly. Holland and Rubin, for example, have made up a motto that expresses this quite succinctly: "No causation without manipulation" (Holland, 1986, p. 959). In other words, only "treatments" that can be assigned randomly to any case at will are considered candidates for exhibiting causal effects. All other attributes of cases, such as gender and race, cannot be causes from this perspective. I agree with others (e.g., Bollen, 1989) who take exception to this restrictive conception of causality, despite the intuitive appeal of counterfactual reasoning. Regardless of whether it can be randomly assigned, any attribute that exposes one to differential treatment by one's environment ought to be considered causal.

When Does a Regression Coefficient Have a Causal Interpretation?

Assuming that we could agree on the definition of a cause, perhaps a more pressing question is: When can a regression coefficient be given a causal interpretation? With nonexperimental data, of course, random assignment to the cause is not possible. In lieu of this, several scholars insist that a fundamental requirement for a causal interpretation to be given to the sample estimate of β in $Y = \beta X + \varepsilon$ is that $\text{Cov}(X,\varepsilon) = 0$, or that the equation disturbance, ε, is uncorrelated with the causal variable. This has been referred to variously as the *pseudoisolation assumption* (Bollen 1989), the *causal assumption* (Clogg and Haritou, 1997), or the *orthogonality condition* (Pearl, 1998). Let us see why this important condition is necessary to causal inferences. Suppose, indeed, that you wish to estimate the model $Y = \beta X + \varepsilon$ using sample data and you believe that the association of X with Y is causal, that is, X causes Y. Suppose, however, that, in truth, a latent variable, ξ, affects both X and Y. Hence, the true model is $X = \gamma_1\xi + \upsilon$, with $\text{Cov}(\xi,\upsilon) = 0$, and $Y = \beta X + \gamma_2\xi + \varepsilon'$, where $\text{Cov}(X,\varepsilon') = \text{Cov}(\xi,\varepsilon') = 0$. [We assume that all variables are centered (i.e., deviated from their means), obviating the need for intercept terms.] Notice, then, that ε is really equal to $\gamma_2\xi + \varepsilon'$. Also, note that $\text{Cov}(X,\xi) = \text{Cov}(\xi, \gamma_1\xi + \upsilon) = \gamma_1 V(\xi)$. Thus, $\text{Cov}(X,\varepsilon) = \text{Cov}(X, \gamma_2\xi + \varepsilon') = \gamma_2\ \text{Cov}(X,\xi) = \gamma_1\gamma_2 V(\xi)$. So if $\text{Cov}(X,\varepsilon)$ is zero, this ensures that one or all of γ_1, γ_2, and $V(\xi)$ equal zero; and this means either that ξ is a constant for every case, in which case it has no real influence on X or Y, or that ξ has no influence on X, or that ξ has no influence on Y. In any of these cases, b from the sample regression is a consistent estimator of β (see the chapter appendix for a discussion of consistency). Otherwise, the sample estimator of β is

$$b = \frac{\text{cov}(X,Y)}{v(X)}$$

and the probability limit of b is

$$\text{plim } b = \frac{\text{plim cov}(X,Y)}{\text{plim } v(X)} \text{ (by the Slutsky theorem), which } = \frac{\text{Cov}(X,Y)}{\sigma_x^2}$$

(since sample estimators of variance and covariance—denoted by lowercase "cov" and "v"—are consistent for their population counterparts—denoted by uppercase "Cov" and "V"), where σ_x^2 denotes the population variance of X and

$$\frac{\text{Cov}(X,Y)}{\sigma_x^2} = \frac{\text{Cov}(X, \beta X + \gamma_2\xi + \varepsilon')}{\sigma_x^2}$$

$$= \frac{\beta\sigma^2{}_x + \gamma_2\ \text{Cov}(X,\xi)}{\sigma_x^2}$$

$$= \beta + \frac{\gamma_2\gamma_1\ V(\xi)}{\sigma^2{}_x}.$$

Hence, b is consistent for $\beta + \gamma_2\gamma_1 V(\xi)/\sigma_x^2$, which is, in general, not the same as β. In fact, if β in the true model is really zero, the value of b may mistakenly attribute the impact of ξ on X, represented by γ_1, and the impact of ξ on Y, represented by γ_2, to a causal effect of X on Y. For this reason, the orthogonality condition is necessary for attributing a causal interpretation to b.

Unfortunately, to assume that the orthogonality condition holds is a great leap of faith. Clogg and Haritou (1997) point out that there is no statistical technique, using the data under scrutiny, for determining whether or not the orthogonality condition obtains. So in practice, researchers often add one or more control variables to the model, inferring that the estimate of X's effect in the model with the "proper variables" controlled is unbiased for the "causal effect." In the words of Clogg and Haritou (1997, p. 84): "Partial regression coefficients or analogous quantities are assumed to be the same as causal effects when the right controls (additional predictors) are included in the model." However, adding variables that are *not* causes of Y to the equation can lead to a failure of the orthogonality condition in the expanded model. This can then result in what Clogg and Haritou (1997) call *included-variable bias*. That is, the estimate of X's effect in the expanded model is biased for the causal effect, due to inclusion of an extraneous variable.

Let's see how this works. Suppose that the true causal model for Y is $Y = \beta X + \varepsilon$ and that the orthogonality condition, $\text{Cov}(X,\varepsilon) = 0$, holds. But you estimate $Y = \beta X + \gamma Z + \upsilon$, where Z is a "predictor" of Y but not a causal influence (e.g., as weight is a predictor of height). For this equation to be valid for causal inference, the necessary causal assumption is $\text{Cov}(X,\upsilon) = \text{Cov}(Z,\upsilon) = 0$. Now ε is actually $\gamma Z + \upsilon$ (the disturbance always contains all predictors of Y that are left out of the current equation). So, since $\text{Cov}(X,\varepsilon) = 0$, we have that $\text{Cov}(X, \gamma Z + \upsilon) = \gamma\,\text{Cov}(X,Z) + \text{Cov}(X,\upsilon) = 0$, or that $\text{Cov}(X,\upsilon) = -\gamma\,\text{Cov}(X,Z)$. Provided that neither γ nor $\text{Cov}(X,Z)$ is zero, the orthogonality condition fails for the estimated model. Hence, the estimate of β from that model is biased for the true causal effect.

Recommendations

In light of the foregoing considerations, one might ask whether we should abandon causal language altogether when dealing with nonexperimental data, as has been suggested by some scholars (e.g., Sobel, 1998). Freedman (1997a,b) is especially critical of drawing causal inferences from observational data, since all that can be "discovered," regardless of the statistical candlepower used, is association. Causation has to be assumed into the structure from the beginning. Or, as Freedman (1997b, p. 182) says: "If you want to pull a [causal] rabbit out of the hat, you have to put a rabbit into the hat." In my view, this point is well taken; but it does not preclude using regression for causal inference. What it means, instead, is that *prior knowledge of the causal status of one's regressors* is a prerequisite for endowing regression coefficients with a causal interpretation, as acknowledged by Pearl (1998). That is, concluding that, say, $\beta \neq 0$ in the equation $Y = \beta X + \varepsilon$ doesn't *demonstrate* that X is a cause of Y. But if X *is* a cause of Y, we should find that β is nonzero in this equation, assuming that all relevant confounds have been controlled. That is, a nonzero β is at least *consistent* with

a causal effect of X on Y. It remains for us to marshal theoretical and/or additional empirical—preferably experimental—grounds for attributing to X causal status in its association with Y. In other words, I think it is quite reasonable to talk of regression parameters as "effects" of explanatory variables on the response, provided that there is a flavor of tentativeness to such language.

Perhaps the proper attitude toward causal inference using regression is best expressed in the following quotes. Clogg and Haritou (1997) recommended that researchers routinely run several regressions that include the focus variable plus all possible combinations of potential confounds and assess the stability of the focus variable's effect across all regressions. They then say (p. 110): "The causal questions that social researchers ask are important ones that we ought to try to answer. If they can only be answered in the context of nonexperimental data, then a method that conveys the uncertainty inherent in the enterprise ought to be sought. We believe that the uncertainty in causal assumptions, not the uncertainty in statistical assumptions and certainly not sampling error, is the most important fact of this enterprise."

Sobel's (1998, p. 346) advice is in the same vein: "[s]ociologists might follow the example of epidemiologists. Here, when an association is found in an observational study that might plausibly suggest causation, the findings are treated as preliminary and tentative. The next step, when possible, is to conduct the randomized study that will more definitively answer the causal question of interest."

In sum, causal modeling via regression, using nonexperimental data, can be a useful enterprise provided we bear in mind that several strong assumptions are required to sustain it. First, regardless of the sophistication of our methods, statistical techniques only allow us to examine *associations* among variables. Thus, the most conservative approach to interpreting β in $Y = \beta X + \varepsilon$ is to say that β represents the expected *difference* in Y for those who are 1 unit apart in X. To say that β reflects the expected *change* in Y were we to *increase* X by 1 unit imparts a uniquely causal interpretation to the X–Y association revealed by the regression. Whether such an interpretation is justified requires additional information, in the form of theory and/or experimental work. At the least, we must assume that $\text{Cov}(X,\varepsilon)$ is zero. This means that no other variable, observed or unobserved, confounds the relationship between X and Y, as in the case of ξ above. As no empirical means exists for checking on this assumption, it is an act of faith. At most we will be able to argue that our findings are *consistent* with a causal effect of X on Y. But only the triangulation of various bits of evidence from many sources, over time, can establish this relation with any authority.

DATASETS USED IN THIS VOLUME

Several datasets are used for examples and exercises throughout the book. Ten of the datasets—those needed for the exercises—can be downloaded from the FTP site for this book at *http://www.wiley.com.* The datasets are in the form of raw data files, easily readable by statistical software programs such as SAS, SPSS, and STATA. Also included at the site are full codebooks in MS Word, listing all variable names and their descriptive labels as well as their order on the data records. Two of the datasets

(*students* and *GSS98*, described below) contain missing values that must be imputed by the reader, as instructed in the exercises. All dataset names below in bold face type indicate data that are available for downloading. The following are brief descriptions of the datasets (names of all downloadable data files and associated codebooks are given in parentheses).

National Survey of Families and Households Datasets

The National Survey of Families and Households (NSFH) is a two-wave panel study of a national probability sample of households in the coterminous United States conducted between 1987 and 1994. Wave 1 of the NSFH, completed in 1988, interviewed 13,007 respondents aged 19 and over living in households in the United States. Certain targeted groups were oversampled: cohabitors, recently married couples, minorities, step-parent families, and one-parent families. For respondents who were cohabiting or married, a shorter, self-administered questionnaire was also given to the partner. The NSFH collected considerable demographic and family information as well as data on more sensitive couple topics such as the quality of the relationship and the manner of handling disagreements, including physical aggression. The survey is described in more detail in Sweet et al. (1988). In wave 2, completed in 1994, interviews were conducted with all 10,005 surviving members of the original sample and with the current spouse or cohabiting partner of the primary respondent. Question sets from the first wave were largely duplicated in the second. The six datasets described below are subsets of this survey.

Couples Dataset (couples.dat; couples.doc). This is a 6% random sample of all married and cohabiting couples from wave 1, with an n of 416 couples. The variables reflect various characteristics of the relationship from both partners' perspectives, as well as items tapping depressive symptomatology of the primary respondent.

Kids Dataset (kids.dat; kids.doc). This consists of a sample of 357 parents and their adult offspring from both waves of the NSFH. Information is contained on couples who were married or cohabiting, with a child between the ages of 12 and 18 in the household in 1987–1988, whose child was also interviewed in 1992–1994. Only cases in which the child had experienced sexual intercourse by 1992–1994 and in which the child had answered the items on sexual permissiveness and sexual behavior were included. Variables reflect attitudes, values, and other characteristics of the parents measured in wave 1, as well as sexual attitudes and behavior reported by their adult offspring in wave 2. Further detail is provided in DeMaris (2002a).

Union Disruption Dataset (disrupt.dat; disrupt.doc). These data consist of 1230 married and cohabiting couples in unions of no more than three years' duration at wave 1 who were followed up in wave 2. Primary interest was in the prediction of union disruption by wave 2, based on various characteristics of the relationship reported in wave 1, including intimate violence. This is a subset of the data used for the larger study reported in DeMaris (2000).

Cohabiting Transitions Dataset (cohabtx.dat; cohabtx.doc). This dataset consists of 411 cohabiting couples in wave 1, followed up in wave 2. It was used to examine the predictors of transition to separation or marriage, as opposed to remaining in the unmarried cohabiting state, by wave 2. Wave 1 characteristics of couples used as predictors of transitions were similar to those for the *union disruption dataset*. The full study is reported in DeMaris (2001).

Wave 1 Couples Dataset. These are the 7273 married and cohabiting couples in wave 1 who constitute the original pool of couples from which the longitudinal violence dataset (described below) was culled. Several characteristics of the relationship were measured in wave 1, with a focus on couple disagreements.

Violence Dataset. These data represent 4095 couples in wave 1 who were still intact in wave 2 and who provided information on patterns of intimate violence at both time periods. The response of interest is the *couple violence profile*, a three-category classification of violence patterns. Predictors are characteristics of the relationship as reported in wave 1. The full study is reported in DeMaris et al. (2003).

Datasets from the NVAWS

NVAWS is short for the national survey on Violence and Threats of Violence Against Women and Men in the United States, 1994–1996, collected by Tjaden and Thoennes (1999). The target population for the NVAWS included men and women from all 50 states and the District of Columbia, and includes 8000 men and 8000 women who were 18 years of age and older in 1994. Datasets employed in this book utilize only the women's data. Variables contain information about four types of victimization experienced over the life course: physical assault, sexual assault, stalking, and threats, as well as the mental health sequelae of such experiences. Three datasets are subsets of this survey.

Victims Dataset. This consists of the 1779 women who reported being victimized at least once by physical or sexual assault, stalking, or intimidation.

Current-Partner Victims Dataset. This is the subset of 331 women from the victims dataset who report being victimized by a current intimate partner.

Minority Women Dataset. These 1343 women are the minority subset of the original 8000 women in the NVAWS.

Other Datasets

Students Dataset (students.dat; students.doc). This is a sample of 235 students taking introductory statistics at Bowling Green State University (BGSU) from the author between the years 1990 and 1999. Variables include student characteristics collected

in the first class session as well as the scores on the first two exams, given, respectively, in the sixth and tenth weeks of the course.

GSS98 Dataset (gensoc.dat; gensoc.doc). These data consist of the 2832 respondents from the 1998 General Social Survey. The GSS is conducted roughly biennially by the National Opinion Research Center. It is based on a multistage probability sample that is representative of all noninstitutionalized English-speaking persons 18 years of age and older living in the household population of the United States. Variables in the dataset represent selected demographic and attitudinal or opinion items deemed by the author to be of interest.

Faculty Salary Dataset (faculty.dat; faculty.doc). This consists of 725 faculty members employed at both the main and Firelands campuses of BGSU during the academic year 1993–1994. Data represent faculty salaries and factors deemed to predict variation in salaries, such as rank and years of seniority. The primary purpose of the study was to discover whether there was any evidence of gender inequity in salary allocation at the institution. Reports of the full studies utilizing these data can be found in Balzer et al. (1996) and Boudreau et al. (1997).

Introductory Sociology Dataset (introsoc.dat; introsoc.doc). These data were taken from all nine sections of introductory sociology offered at BGSU during the 1999 spring semester. The study involved four waves of data collection during the course of the semester. The total sample size is 751 students, but due to absenteeism at one or another data collection point, sample sizes vary in each wave. The focus of the study was an examination of the factors predicting academic performance, particularly self-esteem. Variables consist of measures such as prior and current academic performance, indexes of self-esteem and test anxiety, and related academic factors. Results of the study can be found in Bradley (2000).

Unemployment Transitions Dataset (jobs.dat; jobs.doc). These data are in the form of 620 unemployment spells for 283 Brazilian immigrants residing in the United States and Canada in 1990–1991. The purpose of the study was to test predictions from job search theory regarding the duration in, and rate of exit out of, unemployment for an immigrant population. The predictors consist largely of demographic, familial, and human capital variables. The full study is reported in Goza and DeMaris (2003).

Inmates Dataset (inmates.dat; inmates.doc). This dataset, collected by the Ohio Department of Rehabilitation, consists of information on 1485 male inmates admitted to the Ohio Department of Rehabilitation and Correction during September and October 1985. Variables reflect demographic and criminal history information for each inmate as well as individual lifestyle data and correctional-institution information regarding rule infractions during incarceration. The full study is reported in Clark (2001).

APPENDIX: STATISTICAL REVIEW

Overview

In this appendix I review basic statistical concepts and notation necessary to an understanding of the material in subsequent chapters. I assume that the reader has been exposed to most of this material at a previous time. However, those who are unfamiliar with probability and distribution theory, expectation, variance, covariance, correlation, sampling distributions, parameter estimation, and tests of hypotheses will probably want to read this appendix before proceeding with the rest of the book.

Variables and Their Measurement

The raw material of statistics consists of *data*. Data are essentially measurements for one or more variables, taken on one or more cases, from some population of cases of interest. Let's flesh this idea out a little more. We assume that there is a larger population of cases in which the researcher has an interest. The *population* is simply the collection of cases that the researcher is trying to make general statements about, or "generalize to." *Cases* in the social and behavioral sciences are typically people, but do not have to be. They are the individual units of observation in one's study. These can be individuals or organizations, just as they can be incidents or events. What we typically obtain in sampling cases from the population are attributes or characteristics of the cases, usually expressed as numerical values. These are our *measurements* on the cases. The attributes are called *variables*, and each variable typically exhibits some variability in realized *values* across the n cases in our sample. When the value of a variable for a given case cannot be predicted ahead of time, we refer to that variable as a *random variable*. For example, suppose that I randomly sample a person from the U.S. population and code his or her gender as 1 for male and 0 for female. Then the person's gender is a random variable—I don't know ahead of time what value it will take. If, on the other hand, I divide the population into males and females ahead of time and sample first from the males and second from the females, gender is no longer a random variable. In this case, we say that gender is *fixed*—its value is set ahead of time by the researcher prior to sampling, and there is no mystery about what each case's gender is. This distinction is important in regression modeling when we describe the regressors as random variables versus *fixed effects*.

Variables are distinguished by two major criteria in statistics, both having to do with the specificity of their measurement. The first distinction pertains to *level of measurement*. There are four commonly conceived levels: nominal, ordinal, interval, and ratio. *Nominal variables* are those whose values indicate only qualitative differences in the attribute of interest; they carry no information as to rank order on the attribute. For example, religious affiliation coded 1 for "Protestant," 2 for "Catholic," 3 for "Jewish," and 4 for "other denomination" is a nominal variable. All that can be said about cases with two different values on this attribute is that they are, well, different. Other than that, the numerical codes 1, 2, 3, and 4 convey no quantitative differences on the dimension of religious affiliation.

The values of *ordinal variables*, on the other hand, represent not only qualitative differences but also relative rank order on the attribute. *Religiosity*, for example, coded 1 for "not at all religious," 2 for "slightly religious," 3 for "moderately religious," and 4 for "very religious," is an ordinal variable. Given two people with different religiosity scores, say 3 versus 4, we can say that the second person is "more religious" than the first. How much more religious, however, cannot be specified precisely.

Interval variables represent an even more precise level of measurement. The values of interval variables are distinguished by the fact that they convey the exact amount of the attribute in question. Annual income in dollars, for example, is an interval variable. Further, given two people with different values of income, say $45,529.52 and $51,388.03, we can say not only that their incomes are qualitatively different and that the second person is higher in income but can also specify *precisely how much difference there is in their incomes*: $5858.51, to be exact. Notice, however, that if we collapse income categories into ranges, the variable loses its interval-level specificity and becomes ordinal. For example, suppose that we have income categories defined in $10,000 ranges and coded from 1 for [0–10,000) to 11 for [100,000 or more). Further, suppose that individual A is in category 5 [40,000–50,000) and individual B is in category 6 [50,000–60,000). Certainly, we can say that B has a higher income than A. But it is no longer possible to specify precisely how much higher B's income is.

Ratio variables are interval-level variables with a meaningful zero point. In this case, it makes sense to speak of the ratio of two values. Income is also an example of a ratio variable. If A makes $50,000 a year and B makes $100,000, B makes twice as much income as A.

The other major criterion for distinguishing variables is whether they are discrete or continuous. This distinction is central to the characterization of their probability distributions (see below). Technically, a *discrete variable* is one with a countable number of values. This is a technical concept which essentially means that the values have a one-to-one relationship with the collection of positive integers. Since there are an infinite number of positive integers, discrete variables could conceivably have an infinite number of values. In practice, discrete variables take on only a relatively few values. For example, the number of children ever borne by U.S. women is a discrete variable, taking on values 0, 1, 2, and so on, up to some maximum value delimited by biological possibility, say 25 or so. Nominal variables are always discrete, as are ordinal variables, since rank order can always be put in a one-to-one correspondence with positive integers.

Continuous variables are those with an uncountable number of values. These variables can, technically, take on any value in the real numbers, delimited only by their logical range. Realistically, measurement limitations prevent us from ever actually observing continuous variables in practice. For example, the weight of humans in pounds could conceivably take on any of an uncountably infinite number of values in the range [0–1000]. But limitations in instruments for weight measurement mean that we probably cannot discern weight differences smaller than, say, .001 pound between two people. No matter. We will find it expedient to *treat* variables as

continuous if they are at least ordinal in nature, if they have a sufficient number of values, and if their probability distributions are not too skewed. Otherwise, they will be treated as discrete. For this book, therefore, the discrete–continuous distinction is the one that is most important.

Probability and Distribution Theory

In sampling cases from a population, we speak of the *probability* of observing a particular value for a given variable, for the *i*th individual, where *i* equals 1, 2, . . . , *n*. The technical definition of probability is quite arcane (see, e.g., Chung, 1974; Hoel et al., 1971). Intuitively, however, the probability of some outcome refers to the *relative frequency of its occurrence over an infinite repetition of the conditions that made its observation possible*. For example, if we toss an honest coin, the probability of observing a head is .5. This means that if we were to toss that coin an infinite number of times, 50% of the outcomes would be heads. Since we will never be able to conduct an infinite repetition of any experiment, probabilities are figured by a simple rule. For any event, *E*, the probability of event *E*, or P(*E*), is defined as follows:

$$\mathrm{P}(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{total number of observable outcomes}}$$

Hence, in the coin example, there is only one way to get a head, but there are two possible outcomes of a coin toss: a head or a tail. The probability of a head is therefore $\frac{1}{2} = .5$.

Although in this book we will not be concerned with probability problems per se, a few probability rules are important. First, for any event *A*, if P(*A*) is the probability that *A* occurs, then $1 - \mathrm{P}(A)$ is the probability that it doesn't occur (or that anything else occurs that isn't *A*). Further, consider any two events, *A* and *B*. Then the event (*A* and *B*), also denoted $(A \cap B)$, refers to an event that is both *A* and *B* simultaneously, while the event (*A* or *B*), also denoted $(A \cup B)$, refers to the event that at least one of *A* or *B* occurs. For example, if *A* is "being married" and *B* is "having a child," (*A* and *B*) is "being married with a child," while (*A* or *B*) is satisfied by any of these three events: being married but childless, having a child outside marriage, or being married with a child. The *conditional probability* of an event is the probability of an event under the restriction that some condition holds first. The conditional probability of some event *B*, given that event *A* holds, is denoted $\mathrm{P}(B|A)$. For example, the conditional probability of *B* given *A*, from above, is the conditional probability of having a child given that the person is married. Two events are *independent* if $\mathrm{P}(B) = \mathrm{P}(B|A)$, and *dependent* otherwise. For example, the events "being married" and "having a child" are independent if the probability of having a child is unchanged by whether or not a subject is known to be married. In all likelihood, these events are not independent, since the probability of having a child when one is married is probably higher than the probability of having a child in general, called the *unconditional probability* of having a child. If *A* and *B* are independent events,

P(A and B) = P(A)P(B). This generalizes to: If events A_i are independent, for $i = 1, 2, \ldots, n$, then P(A_1 and A_2 and $\cdots$ and A_n) = P(A_1)P(A_2) $\cdots$ P(A_n).

Probability Distributions. More important for the current work are probability distributions. (Readers with a limited math background may want to review Appendix A, Section I, before proceeding with this section.) A probability distribution for a random variable X is an enumeration of all possible values of X, along with the probability associated with each value, should one collect one observation on X from the population. Actually, this is too simple. In truth, we need to distinguish between the *distribution* and *density* functions for the variable X. The distribution function for X, denoted F(x), tells us P($X \leq x$) for any value x of X. That is, the distribution function tells us the probability of observing any value *up to and including* x, when we make a single observation on X from the population. (I follow the statistical convention here of using X to denote the variable generally and x to denote a specific value of the variable, e.g., 3.2, 5.93, etc.)

What the *density function* tells us, on the other hand, depends on whether X is discrete or continuous. If discrete, the density of x, denoted f(x), gives us the *probability* of getting the specific value x of X when we sample one value of X from the population. Figure 1.1 depicts a simple discrete density function for a variable X.

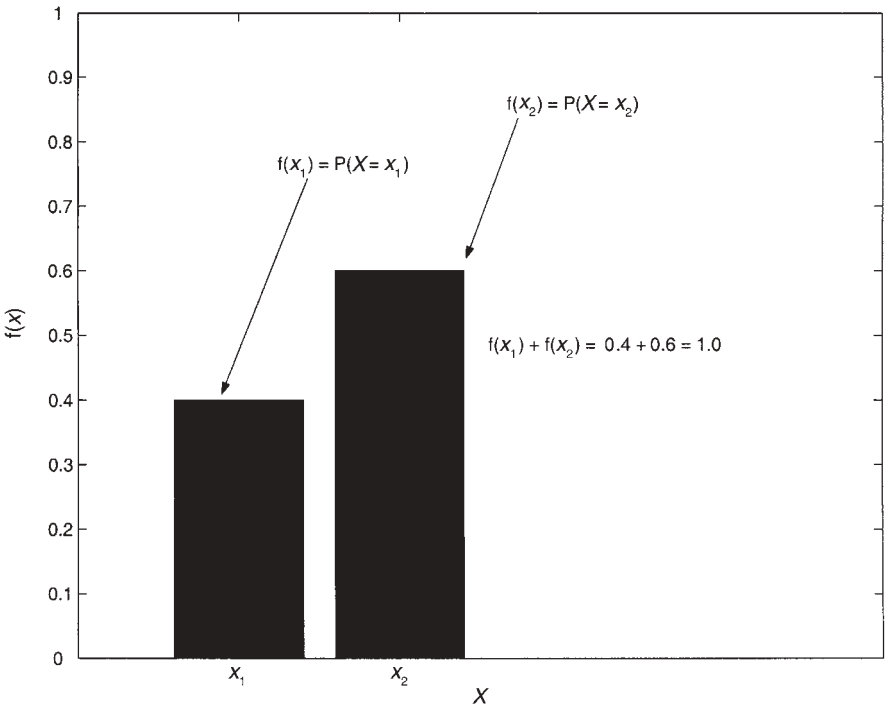

Figure 1.1 Discrete density function for X.