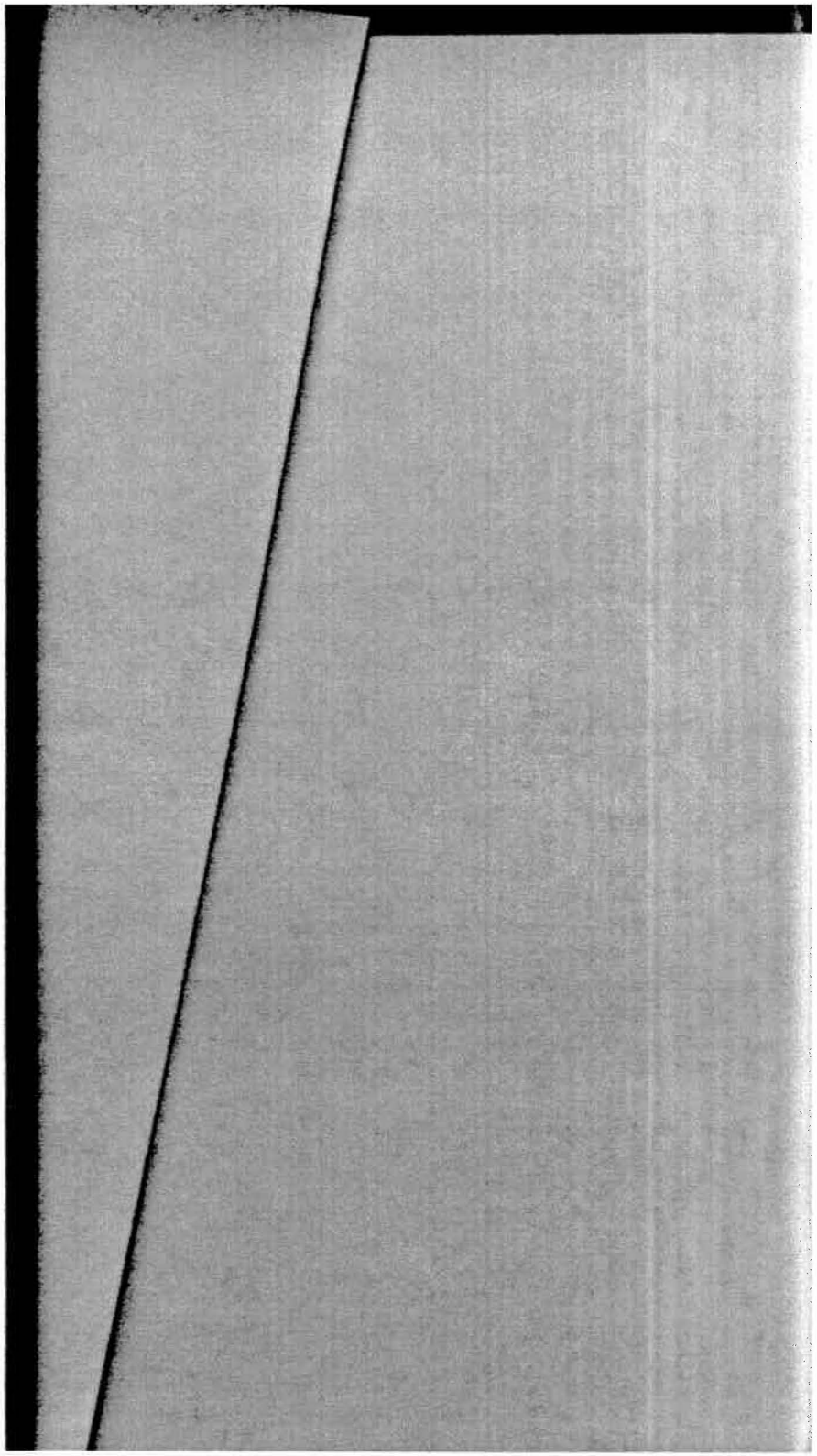
Regression Models for Categorical and Limited Dependent Variables

J. Scott Long

Advanced Quantitative Techniques in the Social Sciences Series 7



Advanced Quantitative Techniques in the Social Sciences

LIST OF ADVISORY BOARD MEMBERS

VOLUMES IN THE SERIES

1. **HIERARCHICAL LINEAR MODELS: Applications and Data Analysis Methods**
   Anthony S. Bryk and Stephen W. Raudenbush
2. **MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Theory**
   John P. Van de Geer
3. **MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Applications**
   John P. Van de Geer
4. **STATISTICAL MODELS FOR ORDINAL VARIABLES**
   Clifford C. Clogg and Edward S. Shihadeh
5. **FACET THEORY: Form and Content**
   Ingwer Borg and Samuel Shye
6. **LATENT CLASS AND DISCRETE LATENT TRAIT MODELS: Similarities and Differences**
   Ton Heinen
7. **REGRESSION MODELS FOR CATEGORICAL AND LIMITED DEPENDENT VARIABLES**
   J. Scott Long
8. **LOG-LINEAR MODELS FOR EVENT HISTORIES**
   Jeroen K. Vermunt
9. **MULTIVARIATE TAXOMETRIC PROCEDURES: Distinguishing Types From Continua**
   Niels G. Waller and Paul E. Meehl
10. **STRUCTURAL EQUATION MODELING: Foundations and Extensions**
    David Kaplan

Regression Models for Categorical and Limited Dependent Variables

J. Scott Long

Advanced Quantitative Techniques in the Social Sciences Series **7**

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Long, J. Scott
Regression models for categorical and limited dependent variables / author, J. Scott Long.

p. cm. — (Advanced quantitative techniques in the social sciences ; v. 7)
Includes bibliographical references and index.
ISBN 0-8039-7374-8 (cloth : alk. paper)
1. Regression analysis. I. Title. II. Series: Advanced quantitative techniques in the social sciences ; 7.
QA278.2.L65 1997 96-35710
519.5'36—dc20

04 05 10 9 8

Production Coordinator: Astrid Virding
Cover Design: Lesa Valdez
Book Design: Ravi Balasuriyia
Print Buyer: Anna Chin

To Valerie and Megan

Contents

List of Figures

List of Tables

Series Editor's Introduction

The tools broadly labeled as "regression" have expanded in number and power over the past two decades. In the "old days," researchers trying to link a set of explanatory variables to a single response variable were essentially limited to the general linear model: analysis of variance—analysis of covariance and multiple regression. These were useful tools when the response variable was measured on an equal interval scale. However, in the social and biomedical sciences, few of the response variables of interest come in equal interval metrics. Responses to survey questions are often, even typically, categorical (e.g., "employed," "unemployed") or ordinal (e.g., "agree," "uncertain," "disagree"). The same holds for the outcomes of people processing and medical institutions: sick or well, arrested or not, dropped out of school or not, lived or died, high school diploma or college degree or postgraduate degree, and so on. For these kinds of response variables, the general linear model is inappropriate and will often give misleading answers.

The solution within a regression framework is "regression-like" models, sometimes collected within the framework of the generalized linear model. The basic idea is still work with a linear combination of explanatory variables, but to allow them to be related to the response variables in a nonlinear way through a "link" function. Then the disturbance is

given an appropriate distribution, usually not the normal. For example, in logistic regression the log of the odds of some binary outcome (e.g., lived or died) is regressed on the usual linear combination of explanatory variables with the underlying conditional distribution of the binary outcomes taken to be binomial.

In this volume, Scott Long addresses these and related kinds of statistical procedures. I am very pleased to add Scott Long's *Limited Dependent Variables* to the series. The topics are of both practical and theoretical importance, and Professor Long has done a excellent job of exposition. The book is well suited as a text for graduate students in the social and biomedical sciences. It will also serve as a wonderful reference for practitioners.

The core of Professor Long's approach is "statistical modeling." A "model" is a simplified rendering of the processes being studied and/or an algebraic representation of a scientific theory. A model is not merely a data reduction device. Given the emphasis on modeling, it is especially important that the techniques discussed be used judiciously and that Professor Long's caveats be taken to heart. Thus, even a state-of-the-art statistical analysis is unlikely to salvage much of use from a seriously flawed dataset. In addition, one must be able to make the case that the statistical model maps well onto the empirical phenomena being studied. Also, researchers use cause-and-effect language at their peril unless there has been real manipulation of the explanatory variables. Finally, when statistical inference is to be undertaken, the sources of uncertainty have to be articulated in a fashion that is consistent with what the model assumes about how the uncertainty operates.

There is really no argument about the validity of these principles, but there are strong disagreements about what these principles mean in practice. To put it a bit (but only a bit) too starkly, at one extreme there are those who never saw a model they did not like. At the opposite extreme are those who never saw a model they liked. Most researchers fall between these extremes where the issues often boil down to where the burden of proof lies—for some, a model is acceptable as long as there is no strong evidence to undermine it. For others, a model is unacceptable unless there is strong evidence to support it. I suspect that social and biomedical researchers tend to fall in the first camp and that statisticians tend to fall more in the second camp. However, from this tension in part has come a range of diagnostic tools that can help (but only help) to determine how sound a model is. Professor Long is to be commended for including a healthy dose of those diagnostics in this book. Practitioners should take them very seriously.

Finally, a word about software. For most of the procedures discussed in those book there exist statistical routines in all of the major statistical packages. This is both a blessing and a curse. The blessing is that minimal computer skills are required. The curse is that minimal computer skills are required. Right answers *and* wrong answers are easy to obtain. With this in mind, Professor Long discusses some of the most popular software. This too deserves serious study.

RICHARD BERK

Preface

This book is about regression models that are appropriate when the dependent variable is binary, ordinal, nominal, censored, truncated, or counted. I refer to these outcomes as categorical and limited dependent variables (CLDVs, for short). Within the last decade, advances in statistical software and increases in computing power have made it nearly as easy to estimate models for CLDVs as the linear regression model. This is reflected in the rapidly increasing use of these models. Nearly every issue of major journals in the social sciences contains examples of models such as logit, probit, or negative binomial regression. While computational problems have largely been eliminated, the models are more difficult to learn and to use. There are two quite different reasons for this. First, the models are nonlinear. As readers will learn well, the nonlinearity of many models for CLDVs makes interpretation of the results more difficult. With the linear regression model, most of the work is done when the estimates are obtained. With models for CLDVs, the task of interpretation is just beginning. Unfortunately, all too often when these models are used, the substantive meaning of the parameters is incompletely explained, incorrectly explained, or simply ignored. Sometimes only the statistical significance or possibly the sign is mentioned. A second reason that these models are difficult to learn is that while models for CLDVs are more complicated than the linear regression model, most

books only discuss them briefly, if at all. While hundreds of pages may be devoted to the linear regression model, only a dozen or two pages are devoted to models for CLDVs.

My goal in writing this book is to provide a unified treatment of the most useful models for categorical and limited dependent variables. Throughout the book, the links among the models are made explicit, and common methods of derivation, interpretation, and testing are applied. Whenever possible, I relate these models to the more familiar linear regression model. While Chapter 2 is a brief review of this model, I assume that readers are familiar with the specification, estimation, and interpretation of the linear regression model.

The best way to learn these models is by seeing them applied to real data and by applying them as you read. To that end, I illustrate each model with data from a variety of applications ranging from attitudes toward working mothers to scientific productivity. You may find it useful to reproduce the results presented in the book using your statistical package. To that end, I have placed the data from the book along with sample programs on my homepage on the World Wide Web (http://www.indiana.edu/~jsl650) or access the Sage Website http://www.sagepub.com/sagepage/authors.HTM for information. While I used GAUSS-Markov for most of the computations, I will be adding sample programs written in Stata, SAS, and LIMDEP. And, a book on using Stata to estimate models for CLDVs is planned.

This book grew out of a course on categorical data analysis taught from 1978 to 1989 at Washington State University and at Indiana University since 1989. Teaching this course is a constant challenge and source of satisfaction. If you find the explanations that follow to be clear, it is largely the fault of those students who refused to accept unclear explanations. (A few refused to accept clear explanations, but that is a different issue.) Questions from students continually motivated me to find a way to make difficult topics accessible. And, indeed, some of the topics are difficult. While I have sought to present the models fully and clearly with the simplest mathematics possible, some readers will find the mathematics to be a challenge. I hope that these readers will persist, because I have yet to find an person who could not master these techniques and use them effectively to learn more about the social world.

J. SCOTT LONG
BLOOMINGTON, IN
JUNE 12, 1996

Acknowledgments

I am indebted to the many people who gave me comments on earlier drafts: Dick Berk, Ken Bollen, Brian Driscoll, Scott Eliason, Lowell Hargens, David James, Bob Kaufman, Herb Smith, Adrian Raftery, Ron Schoenberg, and Yu Xie. Members of the Workshop in Quantitative Methods at Indiana University—Clem Brooks, Bob Carini, Brian Driscoll, Laurie Ervin, David James, Patricia McManus, and Karl Schuessler—gave me feedback that substantially improved the book. Paul Allison, Laurie Ervin, Jacques Hagenaars, Scott Hershberger, and Pravin Trivedi gave me exceptionally detailed and useful advice. Technical Typesetting Inc. did an outstanding job typesetting the book. And, I want to thank C. Deborah Laughton, my editor at Sage, for all that she has done for me and this book. While the suggestions that these people made resulted in a much better book, I am responsible (as they say) for any errors that remain. Research support from the College of Arts and Sciences at Indiana University is gratefully acknowledged.

While planning and writing this book I encountered more than the usual number of problems, few of which were related to the book. My wife Valerie and my daughter Megan shared these challenges with me, and to them I dedicate this book.

Abbreviations and Notation

The following abbreviations and notation are used throughout the book. While I have tried to use consistent notation and to avoid using the same symbol for more than one purpose, there are a few exceptions, such as λ being used as the inverse Mills ratio and the logistic distribution.

Abbreviations

BRM:	binary response model.
cdf:	cumulative density function.
CLDVs:	categorical and limited dependent variables.
CLM:	conditional logit model.
IIA:	independence of irrelevant alternatives.
LM test:	Lagrange multiplier test.
LPM:	linear probability model.
LR test:	likelihood ratio test.
LRM:	linear regression model.
ML:	maximum likelihood.
MNLM:	multinomial logit model.
NB:	negative binomial.
NBRM:	negative binomial regression model.
OLS:	ordinary least squares.

ORM:	ordinal regression model.
pdf:	probability density function.
PRM:	Poisson regression model.
ZINB model:	zero-inflated negative binomial model.
ZIP model:	zero-inflated Poisson model.

Notation

$\approx$:	is approximately equal to (e.g., $\pi \approx 22/7$).
$D(M)$:	the deviance of the model M.
e:	the residual $y - \widehat{y}$.
$\exp(x)$ or e^x:	the exponential of x.
$E(y \mid \mathbf{x}, x_k)$:	the expected value of y given $\mathbf{x}$ and noting the value of x_k.
$f(\cdot)$:	either the logistic pdf $\lambda(\cdot)$ or the normal pdf $\phi(\cdot)$.
$F(\cdot)$:	either the logistic cdf $\Lambda(\cdot)$ or the normal cdf $\Phi(\cdot)$.
$G^2(M_C \mid M_U)$:	the likelihood ratio statistic comparing the constrained model M_C to the unconstrained model M_U.
$G^2(M_\beta)$:	the likelihood ratio statistic comparing M_β to the model with just the intercept or intercepts.
$\mathbf{H}$:	the Hessian matrix of second derivatives of the log likelihood function; also used for the hat matrix in Section 4.2.
i:	the observation number (e.g., x_i).
J:	the number of dependent categories in nominal and ordinal models.
k:	the variable number (e.g., β_k).
K:	the number of x's.
$L(a \mid b)$:	the likelihood of parameters a given data b [e.g., $L(\boldsymbol{\beta} \mid \mathbf{X})$].
LRX^2:	the likelihood ratio chi-square statistic; the same as G^2.
M_C:	the constrained model (i.e., M_U with added constraints).
M_F:	the full model with as many parameters as observations.
M_U:	the unconstrained model.
M_α:	the model with only the intercept or intercepts included.
M_β:	the model with regressors and intercepts included.
N:	the sample size.
$\mathcal{N}(\mu, \sigma^2)$:	the normal distribution with mean μ and variance σ^2.
R^2:	the coefficient of determination.
s^2:	the sample variance of the residual e.
s_k:	the sample standard deviation of x_k.
t:	a t-statistic.
$\text{Var}(\widehat{\boldsymbol{\theta}})$:	the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$.
$\text{Var}(\mathbf{x})$:	the variance-covariance matrix of the x's.
W:	the Wald chi-square test statistic; same as X^2.
X^2:	the Wald chi-square test statistic; the same as W.

x:	the independent variable when there is a single independent variable (e.g., $y = \alpha + \beta x + \varepsilon$).
x_k:	the kth independent variable.
x_k^S:	the kth independent variable standardized to have a variance of 1.
$\overleftarrow{x_k}$:	the lower extreme of x_k; the minimum of x_k if β_k is positive; else the maximum.
$\overrightarrow{x_k}$:	the upper extreme of x_k; the maximum of x_k if β_k is positive; else the minimum.
$\mathbf{x}_i$:	a row vector of independent variables for the ith observation; the ith row of $\mathbf{X}$.
$\bar{\mathbf{x}}$:	a row vector containing the means of the independent variables.
$\mathbf{X}$:	a matrix of independent variables for the entire sample.
y:	the observed dependent variable; in Chapter 7, y is the observed censored variable.
y^*:	the latent dependent variable.
y^S:	y standardized to have a variance of 1.
$y \mid y > \tau$:	the truncated variable y given that y is greater than τ.
z:	a z-statistic.
$\mathbf{z}_{im}$:	a row vector of independent variables for ith observation for outcome m for the CLM in Chapter 7.
α and β:	the intercept and slope when there is a single independent variable (e.g., $y = \alpha + \beta x + \varepsilon$).
α:	the dispersion parameter for the NBRM.
$\boldsymbol{\beta}$:	a vector of coefficients; β_0 is the intercept; β_k is the coefficient for x_k.
β_k:	the unstandardized coefficient for x_k.
$\beta_{k,m\|n}$:	in the MNLM, the coefficient for the effect of x_k on the odds of outcome m versus outcome n.
$\boldsymbol{\beta}_{m\|n}$:	a vector of coefficients $\beta_{k,m\|n}$ in the MNLM.
β_k^S:	the fully standardized coefficient for x_k; y and the x's are standardized.
$\beta_k^{S_x}$:	the x-standardized coefficient for x_k; y is not standardized but x_k is.
$\beta_k^{S_y}$:	the y-standardized coefficient for x_k; y is standardized but the x's are not.
δ:	an abbreviation for $(\mathbf{x}\boldsymbol{\beta} - \tau)/\sigma$ in Chapter 7.
δ_L:	an abbreviation for $(\tau_L - \mathbf{x}\boldsymbol{\beta})/\sigma$ in the two-limit tobit model of Chapter 7.
δ_U:	an abbreviation for $(\tau_U - \mathbf{x}\boldsymbol{\beta})/\sigma$ in the two-limit tobit model of Chapter 7.
$\overline{\Delta}$:	the average absolute discrete change.
$\Delta E(y \mid \mathbf{x})/\Delta x_k$:	the discrete change in y for a change in x_k holding other variables constant.

$\partial E(y \mid \mathbf{x})/\partial x_k$: the partial change in y for an infinitesimal change in x_k holding other variables constant; also called the marginal effect.
ε: the error in equation (e.g., $y^* = \alpha + \beta x + \varepsilon$).
$\boldsymbol{\theta}$: a vector of parameters [e.g., $\boldsymbol{\theta} = (\alpha\, \beta\, \sigma)'$].
$\lambda(\cdot)$, $\Lambda(\cdot)$: the pdf and cdf for the standard logistic distribution with mean 0 and variance $\pi^2/3$.
$\lambda^S(\cdot)$, $\Lambda^S(\cdot)$: the pdf and cdf for the standardized logistic distribution with mean 0 and variance 1.
$\lambda(\cdot)$: the inverse Mills ratio defined as $\phi(\cdot)/\Phi(\cdot)$; used in Chapter 7.
λ_i: the inverse Mills ratio for the ith observation.
μ: the population mean.
$\prod_i y_i$: the product $y_1 \times y_2 \times \cdots$.
σ: the standard deviation of ε given $\mathbf{x}$.
σ_k: the standard deviation of x_k.
σ_y: the standard deviation of y.
τ: the censoring threshold in the tobit, probit, and logit models.
τ_m: the threshold or cutpoint for the ORM.
τ_y: the value assigned to censored cases in tobit models.
τ_L: the lower threshold for the two-limit tobit model.
τ_U: the upper threshold for the two-limit tobit model.
$\phi(\cdot)$, $\Phi(\cdot)$: the pdf and cdf for the standard normal distribution with mean 0 and variance 1.
ψ: the probability of being in a group where the count is always 0. Used with zero modified count models.
$\Omega(\mathbf{x})$: the odds of outcome given $\mathbf{x}$.
$\Omega(\mathbf{x}, x_k)$: the odds of outcome given $\mathbf{x}$ and noting specifically the value of x_k.
$\Omega_m(\mathbf{x})$: the odds of outcomes less than or equal to m versus greater than m.
$\Omega_{m|n}(\mathbf{x})$: the odds of outcome m versus n given $\mathbf{x}$ for the MNLM.

1 Introduction

The linear regression model is the most commonly used statistical method in the social sciences. Hundreds of books describe this model, and thousands of applications can be found. With few exceptions, the regression model assumes that the dependent variable is continuous and has been measured for all cases in the sample. Yet, many outcomes of fundamental interest to social scientists are not continuous or are not observed for all cases. This book considers regression models that are appropriate when the dependent variable is censored, truncated, binary, ordinal, nominal, or count. I refer to these variables as categorical and limited dependent variables (hereafter CLDVs).

A brief review of the literature in the social sciences shows how common CLDVs are. Indeed, continuous dependent variables may be the exception. Here are a few examples:

- *Binary variables* have two categories and are often used to indicate that an event has occurred or that some characteristic is present. Is an adult a member of the labor force? Did a citizen vote in the last election? Does a high school student decide to go to college? Is a consumer more likely to buy the same brand or to try a new brand? Did someone answer a given question on a survey?
- *Ordinal variables* have categories that can be ranked. Surveys often ask respondents to indicate their agreement to a statement using the choices

strongly agree, agree, disagree, and strongly disagree. Items asking the frequency of occurrence might use the categories often, occasionally, seldom, and never. Political orientation may be classified as radical, liberal, and conservative. Educational attainment can be measured in terms of the highest degree received, with the ordinal categories of less than high school, high school, college, and graduate school. Military rank and civil service grade are inherently ordinal.

- *Nominal variables* occur when there are multiple outcomes that cannot be ordered. Occupations can be grouped as manual, trade, blue collar, white collar, and professional. Marital status might be coded as single, married, divorced, and widowed. Political parties in European countries can be considered nominal classifications. Studies of brand preference may include choices among unordered alternatives.

- *Censored variables* occur when the value of a variable is unknown over some range of the variable. The classic example is expenditures for durable goods. Individuals with less disposable income than the price of the cheapest durable good will necessarily have zero expenditure. Measures of workers' hourly wages are restricted on the lower end by the minimum wage rate. Variables measuring percentage, such as the percentage of homes damaged in a natural disaster, are censored below at 0 and above at 100. Censoring can also occur for methodological reasons. In the 1990 Census, all salaries greater than $140,000 were recorded as $140,000 to ensure confidentiality.

- *Count variables* indicate the number of times that some event has occurred. How often did a person visit the doctor last year? How many jobs did someone have? How many strikes occurred? How many articles did a scientist publish? How many political demonstrations occurred? How many children did a family have? How many years of formal education were completed? How many newspapers were founded during a given period?

The level of measurement of a variable is not always clear or unambiguous. Indeed, you might disagree with some of the examples given above. Carter (1971, p. 12) notes that "...statements about levels of measurement of a [variable] cannot be sensibly made in isolation from the theoretical and substantive context in which the [variable] is to be used. Assumptions that a variable is somehow 'intrinsically' interval (ordinal, nominal) are analytically misleading." Education is a good example. Education can be measured as a binary variable that distinguishes those with a high school education or less from others. Or, it could be ordinal indicating the highest degree received: junior high, high school, college, or graduate. Or, it can be a count variable indicating the number of years of school completed. Each of these is reasonable and appropriate depending on the substantive purpose of the analysis.

Once the level of the dependent variable is determined, it is important to match the model used to the level of measurement. If the model chosen assumes the wrong level of measurement, the estimator could be biased, inefficient, or simply inappropriate. Fortunately, there are a large number of models specifically designed for CLDVs. Binary logit and probit are appropriate for binary outcomes. The ordered logit and probit models explicitly deal with the ordered nature of the dependent variable. Multinomial logit is appropriate for nominal outcomes. The tobit model is designed for censored outcomes. Furthermore, a variety of models such as Poisson and negative binomial regression can be used for count outcomes. These and related models are the subject of this book.

Until recently, the greatest obstacle in using models for CLDVs was the lack of software that was flexible, stable, and easy to use. This limitation no longer applies since these models can be estimated routinely with standard software. Now, the greatest impediment is the complexity of the models and the difficulty in interpreting the results. The difficulties arise because most models for CLDVs are nonlinear.

1.1. Linear and Nonlinear Models

The linear regression model is linear, while most models for CLDVs are nonlinear. This difference is so basic for understanding the materials in later chapters that I begin with a general overview of the implications of nonlinearity for interpreting the effects of independent variables. Just as the nonlinearities introduced by relativity theory made physical models substantially more complicated than their Newtonian counterparts, the use of nonlinear statistical models has added new complications for the data analyst.

Figure 1.1 shows a linear and a nonlinear model predicting the dependent variable y. Each model has two independent variables: x is continuous and d is dichotomous with values 0 and 1. To keep the example simple, I assume that there is no random error. Panel A plots the linear model

$$y = \alpha + \beta x + \delta d \qquad [1.1]$$

The solid line beginning at α plots y as x changes when $d = 0$: $y = \alpha + \beta x$. The dashed line beginning at $\alpha + \delta$ plots y as x changes when

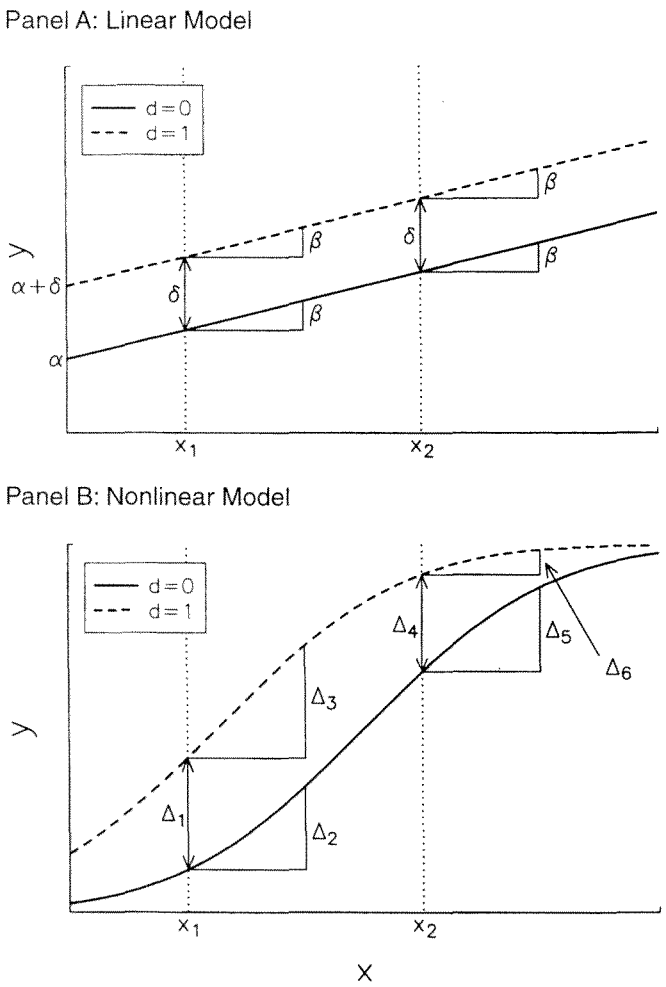


Figure 1.1. Effects of Continuous and Dummy Variables in Linear and Nonlinear Models

$d = 1$: $y = \alpha + \beta x + 1\,\delta = (\alpha + \delta) + \beta x$. The effect of x on y can be computed by taking the partial derivative with respect to x:

$$\frac{\partial y}{\partial x} = \frac{\partial (\alpha + \beta x + \delta d)}{\partial x} = \beta$$

The partial derivative, often called the *marginal effect*, is the ratio of the change in y to the change in x, when the change in x is infinitely small, holding d constant. In a linear model, the partial derivative is the same at *all* values of x and d. Consequently, when x increases by one unit, y increases by β units regardless of the current level of x or d. This is shown in panel A by the four small triangles with bases of length 1 and heights of length β.

The effect of d cannot be computed by taking the partial derivative since d is not continuous. Instead, we measure the *discrete change* in y as d changes from 0 to 1, holding x constant:

$$\frac{\Delta y}{\Delta d} = (\alpha + \beta x + \delta 1) - (\alpha + \beta x + \delta 0) = \delta$$

When d changes from 0 to 1, y changes by δ units regardless of the level of x. This is shown in panel A by the two arrows marking the distance between the solid and dashed lines.

Panel B plots the model

$$y = g(\alpha^* + \beta^* x + \delta^* d) \qquad [1.2]$$

where g is a nonlinear function. For example, for the logit model of Chapter 3, Equation 1.2 becomes

$$y = \frac{\exp(\alpha^* + \beta^* x + \delta^* d)}{1 + \exp(\alpha^* + \beta^* x + \delta^* d)} \qquad [1.3]$$

Interpretation of the effects of x and d is now more complicated. The solid curve for $d = 0$ and the dashed curve for $d = 1$ are no longer parallel: $\Delta_1 \neq \Delta_4$. The effect of a unit change in x differs according to the level of both d and x: $\Delta_2 \neq \Delta_3 \neq \Delta_5 \neq \Delta_6$. The partial derivative of y with respect to x is a function of both x and d. In general, the effect of a unit change in a variable depends on the values of all variables in the model and is no longer simply equal to a parameter of the model.

While Equation 1.2 is nonlinear in y, it is often possible to find some function h that transforms the nonlinear model into a linear model:

$$h(y) = \alpha^* + \beta^* x + \delta^* d$$

For example, we can rewrite Equation 1.3 as

$$\ln\left(\frac{y}{1-y}\right) = \alpha^* + \beta^* x + \delta^* d$$

(*Show this.*[1]) The dependent variable is now $\ln y/(1-y)$, a quantity known as the *logit*. The logit increases by β^* units for every unit increase in x, holding d constant. As with Equation 1.1, this is true regardless of the level of x or d. The problem is that it is often unclear what a unit increase in $h(y)$ means. For example, an increase of β^* in the logit is meaningless to most people.

One of the greatest difficulties in effectively using models for CLDVs is interpreting the nonlinear effects of the independent variables. An all too common, albeit unnecessary, solution is to talk only about the statistical significance of coefficients without indicating how these parameters correspond to meaningful changes in the outcome of interest. A key objective of this book is to show how models for CLDVs can be effectively interpreted.

Throughout the book, I use the term "effect" to refer to a change in an outcome for a change in an independent variable, holding all other variables constant. For example, in the probit model the effect of education on labor force participation might be described as: for an additional year of education the probability of being in the labor force increases by .05, holding all other variables at their means. Or, for count models we might conclude: for each increase in income of $1000, the expected number of children in the family decreases by 5%, holding all other variables constant. The interpretation of an "effect" as causal depends on the nature of the problem being analyzed and the assumptions that a researcher is willing to make. For a detailed discussion of the issues involved in making causal inferences, see Sobel (1995) and the literature cited therein.

1.2. Organization

Chapter 2 reviews the linear regression model to highlight issues that are important for the models in later chapters. Maximum likelihood estimation is introduced within this familiar context to make it is easier to understand how to apply this method to the models in later chapters. Chapter 3 develops models for binary outcomes. I begin with regression of a binary variable to illustrate how CLDVs can cause violations of the assumptions of the linear regression model. Binary probit and logit are first derived using an unobserved or latent dependent variable. I then

[1] Exercises for the reader are given in italics. Solutions are found in the Appendix.

show how the same model can be understood as a nonlinear probability model without appealing to a latent variable. Issues of identification are introduced to explain the apparent differences in results from the logit and probit models. Since numerical methods are often necessary for estimating these models, as well as later models, these methods are discussed in some detail. I also introduce a variety of approaches for interpreting the results from nonlinear models. These techniques are the basis for interpreting all of the models in later chapters. Chapter 4 reviews standard statistical tests associated with maximum likelihood estimation, and considers a variety of measures for assessing the fit of a model. Chapter 5 extends the binary logit and probit models to ordered outcomes. While the resulting ordered logit and probit models are simple extensions of their binary counterparts, having additional outcome categories makes interpretation more complex. Chapter 6 presents the multinomial and conditional logit models for nominal outcomes. The greatest difficulty in using these models is the large number of parameters required and the corresponding problems of interpretation. Chapter 7 considers models with censored and truncated dependent variables, with a focus on the tobit model. The tobit model is developed in terms of a latent variable that is mapped to the observed, censored outcome. The chapter ends by considering a number of related models, including models for sample selection bias. Chapter 8 presents models for count outcomes, beginning with the Poisson regression model. Negative binomial regression and zero modified models are considered as alternatives that allow for overdispersion or heteroscedasticity in the data. Chapter 9 compares and contrasts the models from earlier chapters, and discusses the links between these models and models not discussed in the book, such as log-linear and event history models.

The material in this book can be learned most effectively by reading the chapters in order, but it is possible to skip some chapters or to change the order in which others are read. Everyone should read Chapter 2 to learn the basic terminology and notation. Chapter 3 is essential for all that follows since it introduces key concepts, such as latent variables, and methods of interpretation, such as discrete change. Those who are familiar with Wald and likelihood ratio tests can skip that section of Chapter 4. The discussion of assessing fit in Chapter 4 is not needed for later chapters. Chapter 5 on ordinal outcomes can be read after Chapter 6 on nominal outcomes. Chapter 8 on count models builds on the results for truncated distributions in Chapter 7 to develop the zero modified models. However, most of Chapter 8 is accessible without reading Chapter 7.

While each model studied has unique characteristics, there are important similarities among the models that are exploited. First, each model has the same *systematic component* (McCullagh and Nelder, 1989, pp. 26–27). Specifically, each model enters the independent variables as a linear combination: $\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$. Consequently, in specifying your model you can use all of the "tricks" that you know for entering variables in the linear regression model: nominal variables can be coded as a set of dummy variables; nonlinearities can be introduced by transforming the independent variables; the effects of an independent variable can differ by group by adding interaction variables; and so on. Second, each model is estimated by maximum likelihood. Once the general characteristics of maximum likelihood are understood and the associated statistical tests are learned, these can be applied to all of the models. Third, the same general ideas are used for interpreting each model. Expected values, marginal effects, and discrete changes are computed at interesting values of the independent variables and are presented in plots or tables. Fourth, whenever possible the mathematical tools used for one model are carried over in the presentation of later models.

Many of these models can be derived in different ways. For example, the binary logit model can be developed as a latent variable model in which the observed binary variable is an imperfect measurement of an underlying latent variable. Or, the model can be derived as a discrete choice model in which an individual chooses the outcome that provides the maximum utility. Finally, the model can be viewed as a probability model with the characteristic S-shaped relationship between independent variables and the probability of an event. Each of these approaches results in the same formula relating the independent variables to the expected probability. I show alternative derivations of some models in order to highlight different characteristics of the models. This also serves to link my presentation to the diverse literature in which these models were developed.

Models for CLDVs were often developed independently in different fields, such as biometrics, engineering, statistics, and econometrics, with very little contact across the fields. Consequently, there is no universally accepted notation or terminology. For example, the ordered logit model of Chapter 5 is also known as the ordinal logit model, the proportional odds model, the parallel regression model, and the grouped continuous model. I have tried to use what appears to be emerging as standard terminology within the social sciences. Every effort has been made to keep the notation consistent across chapters. On rare occasions, this has resulted in notation that is different from that commonly used in the literature. To help you keep the notation clear, a table of notation is given on pages xxvii to xxx.

1.3. Orientation

Before ending this chapter, a few words about the general orientation of this book are in order. This is a book about data analysis rather than about statistical theory. The mathematics has been kept as simple as possible without oversimplifying the models in ways that could result in misuse or misunderstanding. The mathematics that is used, however, is essential for understanding the correct *application* of these models. To master the methods, it is important to work with the equations and to try some derivations on your own. To help you do this, I have included exercises in italics at various points. In the long run, it will be worth your while to think about each of these questions before proceeding. Brief answers to the exercises are given in the Appendix.

Seeing how these models can be applied in substantive research is also important for understanding the models. Accordingly, each chapter includes a substantive example that is used to illustrate the interpretation of each model. You are also encouraged to apply these models to your own data while you are reading. To this end, comments are given about four statistical packages for estimating models for CLDVs: LIMDEP Version 7 (Greene, 1995), Markov Version 2 (Long, 1993), SAS Version 6 (SAS Institute, 1990a), and Stata Version 5 (Stata Corporation, 1997). These comments are not designed to teach you how to use these packages, but rather are general comments about difficulties that might be encountered with any statistical package. While nearly all of the analyses in the book were done with my program Markov (Long, 1993) written in GAUSS (Aptech Systems Inc., 1996), any of these four packages could have been used for most analyses. To help you use these methods, I have placed the data sets, programs, and output for the examples on my homepage (http://www.indiana.edu/~jsl650) or access the Sage Website http://www.sagepub.com/sagepage/authors.HTM for information.

While this book contains what I believe are the most basic and useful methods for the analysis of CLDVs, a number of important topics were excluded due to limitations of space. Topics that have not been discussed include: robust and nonparametric methods of estimation, specification tests (Davidson & MacKinnon, 1993, pp. 522–528; Greene, 1993, pp. 648–650), complex sampling, multiple equation systems (see Browne & Arminger, 1995, for a review), and hierarchical models (Longford,

1995, pp. 551–556). Additional citations are given in later chapters. While these are extremely important topics, they presuppose the models considered here and are beyond the scope of this book. I chose a fuller treatment of a smaller number of models rather than less detailed discussion of more methods. Hopefully, this will provide a firm foundation for further reading from the vast and growing literature on limited and categorical dependent variables.

1.4. Bibliographic Notes

Each chapter ends with "Bibliographic Notes." These notes present a brief history of the models in that chapter and provide a list of basic sources.

There are several alternative sources that deal with some of the models presented in this book. Maddala (1983) considers dozens of models for CLDVs. Amemiya (1985) reviews extensions to the tobit model, including sample selection models. McCullagh and Nelder (1989) discuss some of the same models from the standpoint of the generalized linear model. King (1989a) presents many of these models with particular application to political science. Agresti (1990) is particularly useful if all of your variables are nominal or ordinal. Liao (1994) considers the interpretation of probability models within the context of the generalized linear model. Arminger (1995) provides a comprehensive review of many related topics. Finally, Stokes et al. (1995) discuss models for categorical variables in terms of the SAS system.

2 Continuous Outcomes: The Linear Regression Model

This chapter briefly reviews the linear regression model (LRM). While I assume that you are familiar with regression, you should read this chapter carefully since the model is described in a way that facilitates the development of models for categorical and limited dependent variables. Moreover, while the LRM is usually estimated by ordinary least squares, I focus on maximum likelihood estimation since this method is used extensively in later chapters. My discussion of the LRM is by no means comprehensive; for further details, see the references in Section 2.8.

2.1. The Linear Regression Model

The linear regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots + \beta_K x_{iK} + \varepsilon_i \qquad [2.1]$$

where y is the dependent variable, the x's are independent variables, and ε is a stochastic error. The subscript i is the observation number from N random observations. β_1 through β_K are parameters that indicate the effect of a given x on y. β_0 is the intercept which indicates the expected value of y when all of the x's are 0. The model can be written in matrix

notation for all observations as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

If we define $\mathbf{x}_i$ as the ith *row* of $\mathbf{X}$, Equation 2.1 can be written as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

The Assumptions of the LRM

A number of assumptions are added to complete the specification of the model. The first set of assumptions concerns the independent variables.

Linearity. According to Equation 2.1, y is linearly related to the x's through the β parameters. Nonlinear relationships between the x's and y are possible through the inclusion of transformed variables. For example, $y = \beta_0 + \beta_1\sqrt{x_1} + \beta_2 x_1 + \varepsilon$ or $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$. This assumption is considered further in Section 2.4.

Collinearity. The x's are linearly independent. This means that none of the x's is a linear combination of the remaining x's. More formally, this requires that $\mathbf{X}$ is of full rank.

A second set of assumptions concerns the distribution of the error ε, which can be thought of as an intrinsically random, unobservable influence on y. Alternatively, ε can be viewed as the effect of a large number of excluded variables that individually have small effects on y.

Zero Conditional Mean of $\boldsymbol{\varepsilon}$. The conditional expectation of the error is 0:

$$E(\varepsilon_i \mid \mathbf{x}_i) = 0$$

This means that for a given set of values for the x's, the error is expected to be 0. This assumption implies that the conditional expectation of y given $\mathbf{x}$ is a linear combination of the x's:

$$E(y_i \mid \mathbf{x}_i) = E(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta} + E(\varepsilon_i \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

This is shown in Figure 2.1 for the simple regression model: $y = \alpha + \beta x + \varepsilon$. Notice that I use α and β for the parameters in the simple regression model rather than the more cumbersome: $y = \beta_0 + \beta_1 x_1 + \varepsilon$. The expected value of y given x is drawn as a thick line starting at α and moving up and to the right with slope β.

Homoscedastic and Uncorrelated Errors. The errors are assumed to be *homoscedastic*, which means that for a given $\mathbf{x}$, the errors have a constant variance. Formally,

$$\text{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 \quad \text{for all } i$$

When the variance differs across observations, the errors are *heteroscedastic* and $\text{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2$. The errors are also assumed to be uncorrelated across observations, so that for two observations i and j, the covariance between ε_i and ε_j is 0.

In Figure 2.1, the distribution of ε is represented by a dotted curve that should be thought of as coming out of the page into a third dimension.

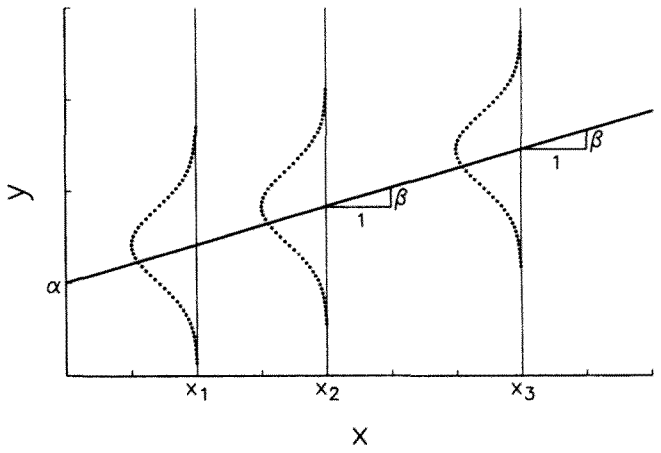


Figure 2.1. Simple Linear Regression Model With the Distribution of y Given x

The higher the curve, the more likely it is to have an error of that value. The errors are homoscedastic since the variance of the error distribution is the same for each x. While the curves are drawn as normal, normality is not required for the errors to be homoscedastic.

Normality. When the errors are thought of as the combined effects of many small factors, it is reasonable to assume that they are normally distributed when conditioned on the x's. With this assumption, the curves in Figure 2.1 should be thought of as normal.

See the references in Section 2.8 for a more detailed discussion of the assumptions.

2.2. Interpreting Regression Coefficients

In Chapter 1, partial derivatives and discrete change were used to describe the effects of an independent variable on the dependent outcome. Even though these two measures of change give identical answers for the LRM, I consider both in order to introduce ideas that are critical in later chapters. The subscript i is dropped to simplify the notation.

The *partial derivative* of y with respect to x_k is

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_k} = \frac{\partial \mathbf{x}\boldsymbol{\beta}}{\partial x_k} = \beta_k$$

In the LRM, the partial derivative is the slope of the line relating y and x_k, holding all other variables constant. Since the model is linear, the value of the partial is a constant β_k that does *not* depend on the level of any of the x's in the model.

The second approach to interpretation involves computing the *discrete change* in the expected value of y for a given change in x_k, holding all other variables constant. The notation $E(y \mid \mathbf{x}, x_k)$ indicates the expected value of y given $\mathbf{x}$, explicitly noting the value of x_k. Thus, $E(y \mid \mathbf{x}, x_k + 1)$ is the expected value of y given $\mathbf{x}$ when the kth variable equals $x_k + 1$. The discrete change in y for a unit change in x_k equals

$$\begin{aligned}\frac{\Delta E(y \mid \mathbf{x})}{\Delta x_k} &= E(y \mid \mathbf{x}, x_k + 1) - E(y \mid \mathbf{x}, x_k) \\ &= [\beta_0 + \beta_1 x_1 + \cdots + \beta_k(x_k + 1) + \cdots + \beta_K x_K + \varepsilon] \\ &\quad - [\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_K x_K + \varepsilon] \\ &= \beta_k\end{aligned}$$

This means that when x_k increases by one unit, y is expected to change by β_k units, holding other x's constant.

In the LRM,

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_k} = \frac{\Delta E(y \mid \mathbf{x})}{\Delta x_k} = \beta_k$$

which allows a simple interpretation of the β's:

- For a unit increase in x_k, the expected change in y equals β_k, holding all other variables constant.

Since dummy variables are coded as 1 if an observation has some characteristic and else 0, the coefficient for a dummy variable can be interpreted in the same way:

- Having characteristic x_k (as opposed to not having the characteristic) results in an expected change of β_k in y, holding all other variables constant.

The slope coefficient is represented in Figure 2.1 by small triangles. The base of each triangle is one unit long, with the rise in the triangle equal to β. Thus, for a unit increase in x, whether starting at x_2, x_3, or any other value of x, y is expected to increase by β units.

2.2.1. Standardized and Semi-Standardized Coefficients

The β coefficients are defined in terms of the original metric of the variables, and are sometimes called *metric coefficients* or *unstandardized coefficients*. It is often useful to compute coefficients after some or all of the variables have been standardized to have a unit variance. This is particularly useful for the models introduced in later chapters where the scale of the dependent variable is arbitrary. This section considers coefficients that are standardized for y, standardized for the x's, and fully standardized for both y and the x's.

y-Standardized Coefficients

Let σ_y be the standard deviation of y. We can standardize y to a variance of 1 by dividing Equation 2.1 by σ_y:

$$\frac{y}{\sigma_y} = \frac{\beta_0}{\sigma_y} + \frac{\beta_1}{\sigma_y}x_1 + \cdots + \frac{\beta_k}{\sigma_y}x_k + \cdots + \frac{\beta_K}{\sigma_y}x_K + \frac{\varepsilon}{\sigma_y}$$

Adding new notation,

$$y^S = \beta_0^{S_y} + \beta_1^{S_y} x_1 + \cdots + \beta_k^{S_y} x_k + \cdots + \beta_K^{S_y} x_K + \varepsilon^{S_y}$$

where y^S is y standardized to have a unit variance. $\beta_k^{S_y} = \beta_k / \sigma_y$ is a *semi-standardized coefficient with respect to y* or simply a *y-standardized coefficient*. It is still the case that

$$\frac{\partial E(y^S \mid \mathbf{x})}{\partial x_k} = \frac{\Delta E(y^S \mid \mathbf{x})}{\Delta x_k} = \beta_k^{S_y}.$$

For a continuous variable, $\beta_k^{S_y}$ can be interpreted as:

- For a unit increase in x_k, y is expected to change by $\beta_k^{S_y}$ standard deviations, holding all other variables constant.

For a dummy variable,

- Having characteristic x_k (as opposed to not having the characteristic) results in an expected change in y of $\beta_k^{S_y}$ standard deviations, holding all other variables constant.

x-Standardized Coefficients

Let σ_k be the standard deviation of x_k. Then, dividing each x_k by σ_k and multiplying the corresponding β_k by σ_k,

$$y = \beta_0 + (\sigma_1 \beta_1)\frac{x_1}{\sigma_1} + \cdots + (\sigma_k \beta_k)\frac{x_k}{\sigma_k} + \cdots + (\sigma_K \beta_K)\frac{x_K}{\sigma_K} + \varepsilon$$

and, adding new notation,

$$y = \beta_0 + \beta_1^{S_x} x_1^S + \cdots + \beta_k^{S_x} x_k^S + \cdots + \beta_K^{S_x} x_K^S + \varepsilon$$

where x_k^S is x_k standardized to have a unit variance, and $\beta_k^{S_x} = \sigma_k \beta_k$ is a *semi-standardized coefficient with respect to x* or simply an *x-standardized coefficient*. For a continuous variable, $\beta_k^{S_x}$ can be interpreted as:

- For a standard deviation increase in x_k, y is expected to change by $\beta_k^{S_x}$ units, holding all other variables constant.

Fully Standardized Coefficients

It is also possible to standardize both y and the x's:

$$\frac{y}{\sigma_y} = \frac{\beta_0}{\sigma_y} + \left(\frac{\sigma_1 \beta_1}{\sigma_y}\right)\frac{x_1}{\sigma_1} + \cdots + \left(\frac{\sigma_k \beta_k}{\sigma_y}\right)\frac{x_k}{\sigma_k} + \cdots + \left(\frac{\sigma_K \beta_K}{\sigma_y}\right)\frac{x_K}{\sigma_K} + \frac{\varepsilon}{\sigma_y}$$

and, adding new notation,

$$y^S = \beta_0^S + \beta_1^S x_1^S + \cdots + \beta_k^S x_k^S + \cdots + \beta_K^S x_K^S + \varepsilon^{S_y}$$

$\beta_k^S = (\sigma_k \beta_k)/\sigma_y$ is a *fully standardized coefficient* or a *path coefficient*. Since

$$\frac{\partial E(y^S \mid \mathbf{x}^S)}{\partial x_k^S} = \frac{\Delta E(y^S \mid \mathbf{x}^S)}{\Delta x_k^S} = \beta_k^S$$

the following interpretation applies:

- For a standard deviation increase in x_k, y is expected to change by β_k^S standard deviations, holding all other variables constant.

Standardized Coefficients for Dummy Variables

For a dummy variable, the meaning of a standard deviation change is unclear. For example, consider the variable *MALE* defined as 1 for men and 0 for women. Assume that the regression coefficient for *MALE* equals .5. The effect of *MALE* changing from 0 to 1 is quite clear: being male increases the dependent variable by .5, holding all other variables constant. Now consider the x-standardized coefficient. Suppose that the standard deviation of *MALE* is .25. Then the x-standardized coefficient would equal .125 (= .5 × .25). To say that a standard deviation change in a person's gender increases the dependent variable by .125 does not make substantive sense. While fully standardized and x-standardized coefficients for dummy variables are sometimes used to compare the magnitudes of the effects of variables, I do not find such comparisons useful. Consequently, x-standardized and fully standardized coefficients for dummy variables are not used in later chapters.

Comparison to Nonlinear Models

The interpretation of the coefficients in the LRM differs in two important respects from the nonlinear models in later chapters. First, in

nonlinear models, $\partial E(\cdot)/\partial x_k$ depends on the value of x_k and on the values of the other x's in the model. Second, in nonlinear models, $\partial E(\cdot)/\partial x_k$ does not necessarily equal $\Delta E(\cdot)/\Delta x_k$. *It is extremely important to avoid generalizing the simple interpretation of the LRM to the models in later chapters.*

2.3. Estimation by Ordinary Least Squares

Ordinary least squares (OLS) is the most frequently used method of estimation for the LRM. The OLS estimator of $\boldsymbol{\beta}$ is that value $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of the squared residuals: $\sum_{i=1}^{N}(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})^2$. The resulting estimator is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with the covariance matrix:

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \begin{pmatrix} \mathrm{Var}(\widehat{\beta}_0) & \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) & \cdots & \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_K) \\ \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_0) & \mathrm{Var}(\widehat{\beta}_1) & \cdots & \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\widehat{\beta}_K, \widehat{\beta}_0) & \mathrm{Cov}(\widehat{\beta}_K, \widehat{\beta}_1) & \cdots & \mathrm{Var}(\widehat{\beta}_K) \end{pmatrix}\end{aligned}$$

When the assumptions of the model hold, the OLS estimator is the best linear unbiased estimator. This means that if the assumptions hold, the OLS estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator [i.e., $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$] that has the minimum variance among all linear estimators.

To estimate $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$, we need an estimate of the variance of the errors, σ^2. Defining the residual as $e_i = y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}$, we can use the unbiased estimator:

$$s^2 = \frac{1}{N-K-1}\sum_{i=1}^{N} e_i^2$$

where K is the number of independent variables. This allows us to estimate the covariance matrix as $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$. If the errors are normal and $\beta_k = \beta^*$, then

$$t_k = \frac{\widehat{\beta}_k - \beta^*}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}}$$

has a t-distribution with $N - K - 1$ degrees of freedom and can be used to test the hypothesis that H_0: $\beta_k = \beta^*$. Without assuming normality, t_k has a t-distribution as the sample becomes infinitely large (Greene, 1993, pp. 299–301). Issues involved in testing hypotheses are discussed in Chapter 4.

Example of the LRM: Prestige of the First Job

Long et al. (1980) examined factors that affect the prestige of a scientist's first academic job for a sample of male biochemists. Their primary interest was whether characteristics associated with scientific productivity were more important than characteristics associated with educational background. Here I extend those analyses to include information on female scientists.

The dependent variable is the prestige of the first job (*JOB*). Prestige is rated on a continuous scale from 1.00 to 5.00, with schools from 1.00 to 1.99 classified as adequate, those from 2.00 to 2.99 as good, 3.00 to 3.99 as strong, and those above 3.99 as distinguished. Graduate programs rated below adequate or departments without graduate programs were coded as 1.00. The implications of this decision are considered in Chapter 7 when this example is used to illustrate the tobit model. The independent variables are described in Table 2.1. Our regression model is

$$JOB = \beta_0 + \beta_1 FEM + \beta_2 PHD + \beta_3 MENT + \beta_4 FEL + \beta_5 ART + \beta_6 CIT + \varepsilon$$

Table 2.2 presents the estimates of the unstandardized and standardized coefficients. t-values are also presented, but are not discussed until

TABLE 2.1 Descriptive Statistics for the First Academic Job Example

Name	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*	*Description*
JOB	2.23	0.97	1.00	4.80	Prestige of job (from 1 to 5)
FEM	0.39	0.49	0.00	1.00	1 if female; 0 if male
PHD	3.20	0.95	1.00	4.80	Prestige of Ph.D. department
MENT	45.47	65.53	0.00	532.00	Citations received by mentor
FEL	0.62	0.49	0.00	1.00	1 if held fellowship; else 0
ART	2.28	2.26	0.00	18.00	Number of articles published
CIT	21.72	33.06	0.00	203.00	Number of citations received

NOTE: $N = 408$.

TABLE 2.2 Linear Regression of the Prestige of the First Academic Job

Name	β	β^{S_x}	β^{S_y}	β^{S}	t
Constant	1.067	—	—	—	6.42
FEM	−0.139	—	−0.143	—	−1.54
PHD	0.273	0.260	0.280	0.267	5.53
MENT	0.001	0.078	0.001	0.080	1.69
FEL	0.234	—	0.240	—	2.47
ART	0.023	0.051	0.023	0.053	0.79
CIT	0.004	0.148	0.005	0.153	2.28

NOTE: $N = 408$. β is an unstandardized coefficient; β^{S_x} is an x-standardized coefficient; β^{S_y} is a y-standardized coefficient; β^{S} is a fully standardized coefficient; t is a t-test of β.

Chapter 4. The variables *FEM* and *CIT* can be used to illustrate the interpretation of coefficients.

- *Unstandardized coefficients*. Being a female scientist decreases the expected prestige of the first job by .14 points on a five-point scale, holding all other variables constant. For every additional citation, the prestige of the first job is expected to increase by .004 units, holding all other variables constant. (This effect is small due to the large standard deviation in *CIT*.)
- *x-standardized coefficients*. For every standard deviation increase in citations, the prestige of the first job is expected to increase by .15 units, holding all other variables constant.
- *y-standardized coefficients*. Being a woman decreases the expected prestige of the first job by .14 standard deviations, holding all other variables constant. For every additional citation, the prestige of the first job is expected to increase by .005 standard deviations, holding all other variables constant. (The unstandardized and *y*-standardized coefficients are nearly identical since the variance of y is about 1.)
- *Fully standardized coefficients*. For every standard deviation increase in citations, the prestige of the first job is expected to increase by .15 standard deviations, holding all other variables constant.

Both fully standardized and *y*-standardized coefficients are used to interpret many of the models in later chapters.

2.4. Nonlinear Linear Regression Models

While the LRM is a linear model, nonlinear relationships between the independent variables and the dependent variable can be incorporated by transforming the variables. For example, consider the nonlinear model:

$$z = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon) \qquad [2.2]$$

If we take the log of both sides,

$$\ln(z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

the resulting equation is linear in $\ln(z)$ even though it is nonlinear in z. Accordingly, the slope β_1 can be interpreted as discussed above: for a unit increase in x_1, $\ln(z)$ is expected to increase by β_1 units, holding x_2 constant. Note, however, that a β_1 unit increase in $\ln(z)$ from 1 to $1 + \beta_1$ involves a different change in z than a change in $\ln(z)$ from, say, 2 to $2 + \beta_1$. This can be seen by taking the derivative of z with respect to x:[1]

$$\begin{aligned}
\frac{\partial z}{\partial x_1} &= \frac{\partial \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\partial x_1} \\
&= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)\frac{\partial(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\partial x_1} \\
&= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)\beta_1 \\
&= z\beta_1
\end{aligned}$$

Thus, even though the expected change in $y = \ln(z)$ is the same regardless of the current levels of x_1 and x_2, the change in z [not $\ln(z)$] depends on the level of z.

Equation 2.2 is an example of a class of nonlinear models known as *log-linear models:* while z is nonlinearly related to the x's, the log of z is linearly related to the x's. Since the logit models of Chapters 3, 4, and 6 and the count models of Chapter 8 are log-linear models, it is worth considering a simple method of interpretation that can be used for any log-linear model.

Since $\exp(a + b) = \exp(a)\exp(b)$, Equation 2.2 can be written as

$$z(x_1) = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\varepsilon)$$

[1] This requires the chain rule:

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)}\frac{\partial g(x)}{\partial x} \quad \text{and} \quad \frac{\partial \exp(x)}{\partial x} = \exp(x).$$

where $z(x_1)$ indicates the value of z when x_1 has a given value. Consider increasing x_1 by 1 to $x_1 + 1$:

$$\begin{aligned} z(x_1 + 1) &= \exp(\beta_0)\exp[\beta_1(x_1 + 1)]\exp(\beta_2 x_2)\exp(\varepsilon) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_1)\exp(\beta_2 x_2)\exp(\varepsilon) \end{aligned}$$

The ratio of $z(x + 1)$ and $z(x)$ is the multiplicative factor change in z for a unit change in x_1:

$$\frac{z(x_1 + 1)}{z(x_1)} = \frac{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_1)\exp(\beta_2 x_2)\exp(\varepsilon)}{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\varepsilon)} = \exp(\beta_1)$$

This leads to the following interpretation:

- For a unit increase in x_1, z is expected to change by the factor $\exp(\beta_1)$, holding all other variables constant.

Or, the percentage change in z for a unit change in x_1 can be computed as

$$100\,\frac{z(x_1 + 1) - z(x_1)}{z(x_1)} = 100\left[\frac{z(x_1 + 1)}{z(x_1)} - \frac{z(x_1)}{z(x_1)}\right] = 100[\exp(\beta_1) - 1]$$

This can be interpreted as:

- For a unit increase in x_1, z is expected to change by $100[\exp(\beta_1) - 1]\%$, holding all other variables constant.

Note that other nonlinear models do not have this simple interpretation in terms of a factor or a percentage change.

2.5. Violations of the Assumptions

While a complete discussion of the consequences of violating the assumptions of the LRM is beyond the scope of my review, I consider two violations that are particularly useful for understanding the models in later chapters.

2.5.1. The Nonzero Conditional Mean of ε

In the LRM,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon \qquad [2.3]$$

we assume that $E(\varepsilon \mid \mathbf{x}) = 0$. Consider a simple modification where we now assume that $E(\varepsilon \mid \mathbf{x}) = \delta$. Here δ is an unknown, *nonzero* constant. We can modify Equation 2.3 so that the new error will have a zero mean:

$$\begin{aligned} y &= (\beta_0 + \delta) + \beta_1 x_1 + \cdots + \beta_K x_K + (\varepsilon - \delta) \\ &= \beta_0^* + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon^* \end{aligned}$$

We have subtracted the mean of ε $(= \delta)$ from ε to create a new error ε^* with a zero mean. (*Show that the mean of* ε^* *is* 0.) To maintain the equality, we also added δ which is combined with β_0 and relabeled as β_0^*. The resulting equation has all of the properties of the LRM, including a mean of 0 for the error ε^*. Consequently, we can use OLS to obtain best, linear, unbiased estimates of β_0^* (not β_0) and the β_k's. The expected value of $\widehat{\beta}_0^*$ is a combination of the intercept β_0 and the mean of ε: $E(\widehat{\beta}_0^*) = \beta_0 + \delta$. No matter how large the sample, it is impossible to disentangle estimates of β_0 and δ. More formally, β_0 and δ are not identified individually, although their sum $\beta_0 + \delta$ is identified.

Since the idea of *identification* is essential for understanding models for CLDVs, it is worth reinforcing the key ideas with Figure 2.2. Assume that the sample data, which are indicated by the dots, are generated by the model $y = \alpha + \beta x + \varepsilon$, where ε is normally distributed with mean δ. The solid line represents $E(y \mid x) = \alpha + \beta x$. As would be expected, the

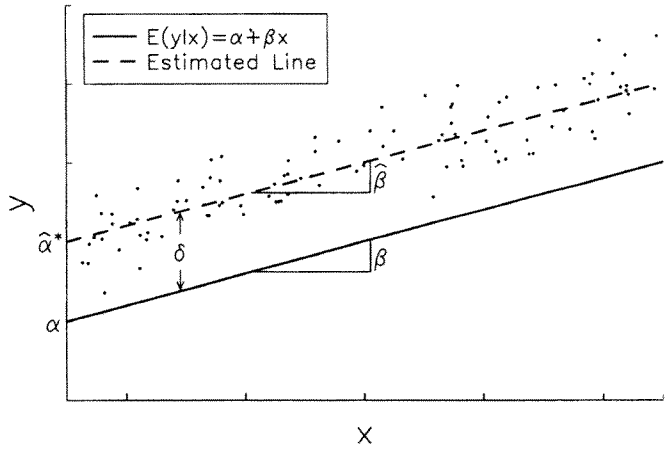


Figure 2.2. Identification of the Intercept in the Linear Regression Model

observed data are located approximately $\delta = E(\varepsilon \mid x)$ units above the regression line. The OLS *estimate* of the regression line is the dashed line that runs through the observations, with intercept $\widehat{\alpha}^*$ and slope $\widehat{\beta}$. The sample estimate of the slope appears unaffected by the nonzero mean of the errors, and is approximately equal to β. Consistent with our algebraic argument, the estimated intercept is about δ units above the population intercept α as a consequence of the nonzero mean of the errors. While neither α nor δ is identified, the sum $\alpha + \delta$ is identified and can be estimated by $\widehat{\alpha}^*$.

This simple example illustrates a number of critical ideas related to the concept of identification. First, a parameter is unidentified when it is impossible to estimate a parameter regardless of the data available. Identification is a limitation of the model that cannot be remedied by increasing the sample size. Second, models become identified by adding assumptions. The intercept β_0 is identified if we assume that $E(\varepsilon \mid \mathbf{x}) = 0$; without this assumption it is unidentified. Third, it is possible for some parameters to be identified while others are not. Thus, while β_0 is not identified unless the value of $E(\varepsilon \mid \mathbf{x})$ is assumed, β_1 through β_K are identified without this assumption. Finally, while individual parameters may not be identified, combinations of those parameters may be identified. Thus, while neither δ nor β_0 is identified, the sum $\beta_0 + \delta$ is identified. These ideas are important for understanding how we identify the models in later chapters.

2.5.2. The x's and ε Are Correlated

The assumption $E(\varepsilon \mid \mathbf{x}) = 0$ implies that the x's and ε are uncorrelated. In practice, there are several reasons why the x's might be correlated with the errors, including reciprocal effects among variables, measurement error, incorrect functional form, and β's that differ across observations (Kmenta, 1986, pp. 334–350). Here I consider the effect of excluding a variable since this will help us understand the tobit model in Chapter 7.

If we estimate a model that excludes an independent variable which is correlated with included independent variables, the OLS estimates are biased and inconsistent. Kmenta (1986, pp. 443–446) shows that this is due to the correlation between the error and the independent variables in the misspecified model. To see why they are correlated, assume that the data are generated by the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad [2.4]$$

but that we have estimated the model:

$$y = \beta_0 + \beta_1 x_1 + \nu \qquad [2.5]$$

The error ν absorbs the excluded variable x_2 and the original ε:

$$\nu = \beta_2 x_2 + \varepsilon$$

If x_1 and x_2 are correlated, then ν and x_1 must be correlated. (*Why must this be the case?*) Consequently, the OLS estimate of β_1 in Equation 2.5 is a biased and inconsistent estimate of β_1 in Equation 2.4.

2.6. Maximum Likelihood Estimation

If we assume that the errors are normally distributed, the LRM can be estimated by maximum likelihood (ML). While the OLS and ML estimators of $\boldsymbol{\beta}$ are identical for the LRM, I introduce ML estimation within the familiar context of regression to make it easier to understand the application of ML to the models in later chapters.

2.6.1. Introduction to ML Estimation

Consider the problem of estimating the probability of having a given number of men in your sample. The binomial formula computes the probability of having s men in a sample of size N with the population parameter π indicating the probability of being male:

$$\Pr(s \mid \pi, N) = \frac{N!}{s!(N-s)!}\,\pi^s(1-\pi)^{N-s} \qquad [2.6]$$

where $k! = k \cdot (k-1) \cdots 2 \cdot 1$. For example, the probability of having three men in a sample of 10 with the probability of being a male equal to .5 is

$$\Pr(s = 3 \mid \pi = .5,\ N = 10) = \frac{10!}{3!7!}\,.5^3(1-.5)^7 = 0.117$$

This is a typical problem in probability. We know the formula for the probability distribution and the values of the parameters π and N. We want to know the probability of a particular outcome s. In statistics, we know s and N, but want to estimate π from the sample information. *The ML estimate is that value of the parameter that makes the observed data most likely.*

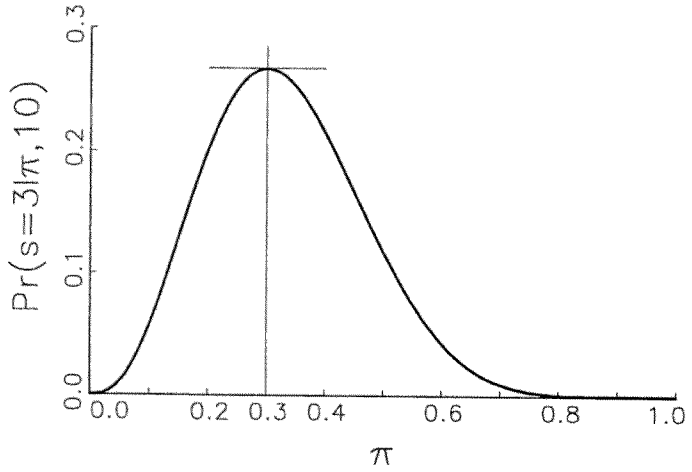


Figure 2.3. Probability of $s = 3$ for Different Values of π

To continue our example, assume that we know that $s = 3$ and $N = 10$, but that we do not know π. What value of π is most likely to have generated the observed $s = 3$? Figure 2.3 plots the probability of observing three successes out of 10 tries for all possible values of π. The tangent on the top of the curve shows that the highest probability occurs at .3. Thus, $\widehat{\pi} = .3$ is our ML estimate.

2.6.2. The Likelihood Function

When Equation 2.6 is thought of as computing the probability of s events as a function of the parameters N and π, it is referred to as a probability function: the values of N and π are held constant while s varies. When we think of the same equation as a function of π, we refer to it as a *likelihood function*: the values of N and s are held constant while π varies. The likelihood function for our example is

$$L(\pi \mid s = 3,\ N = 10) = \frac{10!}{3!7!}\pi^3(1-\pi)^7$$

The maximum likelihood estimate is that value $\widehat{\pi}$ that maximizes the likelihood of observing the sample data that were actually observed. The maximum occurs when the derivative of the likelihood function, called

the *gradient* or *score*, equals 0:

$$\frac{\partial L(\pi \mid s = 3, N = 10)}{\partial \pi} = 0$$

This is represented in Figure 2.3 by the tangent line with slope 0 located at $\pi = .3$.

The value that maximizes the likelihood function also maximizes the log of the likelihood. Since it is generally easier to solve the gradient of the log likelihood than the likelihood itself, the ML estimate is usually computed by solving the equation:

$$\frac{\partial \ln L(\pi \mid s = 3,\ N = 10)}{\partial \pi} = 0$$

For our example,[2]

$$\begin{aligned}
\frac{\partial \ln L(\pi \mid s = 3,\ N = 10)}{\partial \pi} &= \frac{\partial \ln[(10!/3!7!)\pi^3(1-\pi)^7]}{\partial \pi} \\
&= \frac{\partial \ln(10!/3!7!)}{\partial \pi} + \frac{\partial 3\ln \pi}{\partial \pi} + \frac{\partial 7\ln(1-\pi)}{\partial \pi} \\
&= 0 + \frac{\partial 3\ln \pi}{\partial \pi} + \frac{\partial 7\ln(1-\pi)}{\partial(1-\pi)}\frac{\partial(1-\pi)}{\partial \pi} \\
&= \frac{3}{\pi} - \frac{7}{1-\pi}
\end{aligned}$$

Setting $\partial \ln L(\pi \mid s = 3, N = 10)/\partial \pi = 0$ and solving for π results in $\widehat{\pi} = .3 = s/N$.

2.6.3. ML Estimation of the Sample Mean

Before estimating the regression model with ML, it is useful to consider the similar but simpler problem of estimating the mean of a standard normal distribution. If y is drawn from a normal distribution with a standard deviation of 1, then the probability density function (pdf) for y is

$$f(y_i \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(y_i-\mu)^2}{2}\right)$$

[2] We use the chain rule:

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)}\frac{\partial g(x)}{\partial x} \quad \text{and} \quad \frac{\partial \ln x}{\partial x} = \frac{1}{x}$$

Since μ is unknown, we write the likelihood function as

$$L(\mu \mid y_i,\ \sigma = 1) = f(y_i \mid \mu,\ \sigma = 1)$$

For three independent observations, the likelihood is the product of the individual likelihoods:

$$L(\mu \mid \mathbf{y},\ \sigma = 1) = \prod_{i=1}^{3} L(\mu \mid y_i,\ \sigma = 1) = \prod_{i=1}^{3} f(y_i \mid \mu,\ \sigma = 1)$$

and the log likelihood is

$$\ln L(\mu \mid \mathbf{y},\ \sigma = 1) = \sum_{i=1}^{3} \ln L(\mu \mid y_i,\ \sigma = 1) = \sum_{i=1}^{3} \ln f(y_i \mid \mu,\ \sigma = 1)$$

The ML estimate is the value $\widehat{\mu}$ that maximizes this equation.

To get a better sense of how the ML estimate is determined, consider Figure 2.4. Suppose that there are three observations with values

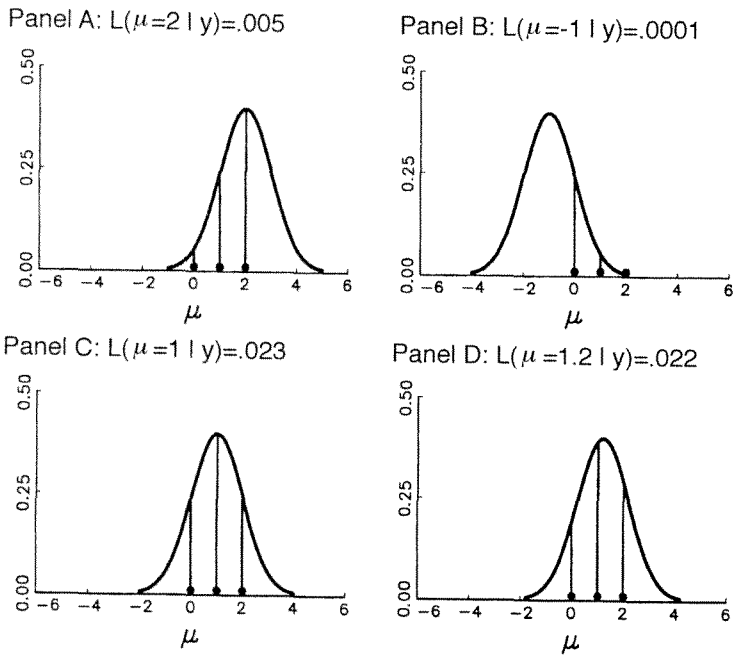


Figure 2.4. Maximum Likelihood Estimation of μ From a Normal Distribution

0, 1, and 2. These are represented in the figure as solid circles. The four panels correspond to a sequence of guesses for the value of μ that maximizes the likelihood. In panel A, the normal curve is centered on $\mu = 2$. The likelihood of each point is indicated by a vertical line, with the overall likelihood equal to the product of the lengths of the lines: $L(\mu = 2 \mid \mathbf{y}) = .005$. Panel B computes the likelihood for $\mu = -1$, resulting in $L(\mu = -1 \mid \mathbf{y}) = .0001$. To increase the likelihood, we need a value of μ somewhere between 2 and -1. Panel C shows $\mu = 1$, resulting in $L(\mu = 1 \mid \mathbf{y}) = .023$. When we increase the mean slightly to 1.2 in panel D, the likelihood is reduced to .022. Of our four tries, $\mu = 1$ produces the largest likelihood. Tentatively, we conclude that $\widehat{\mu}_{\text{ML}} = 1$.

In practice, ML is more complicated. First, we would usually have more observations. Second, we would often be estimating more than one parameter (e.g., μ and σ). Finally, we would have to consider all possible values of the parameters being estimated, not just the four values in our figure. Still, the general ideas are the same.

2.6.4. ML Estimation for Regression

Maximum likelihood for the LRM is a direct extension of fitting a normal distribution to a set of points. Consider estimating the simple regression $y = \alpha + \beta x + \varepsilon$ using three observations: (x_1, y_1), (x_2, y_2), and (x_3, y_3). Panels A and B of Figure 2.5 compare the likelihoods for two sets of possible estimates. The observed data are indicated by circles. The assumed distribution of y conditional on x is represented by the normal curves which should be visualized as coming out of the page into a third dimension. The likelihood of an observation for a given pair α and β is indicated by the length of the line from an observation, indicated by a circle, to the normal curve. In panel A for α^a and β^a, we find that (x_3, y_3) is very unlikely, while (x_1, y_1) is quite likely. The likelihood of α^a and β^a is the product of the three lines in panel A. Clearly, α^a and β^a are not the ML estimates since it is easy to find other estimates that increase the likelihood, such as α^b and β^b in panel B. The ML estimates are those values $\widehat{\alpha}$ and $\widehat{\beta}$ that make the likelihood as large as possible.

Mathematically, we can develop the ML estimator for the LRM as follows. Since y conditional on x is distributed normally with mean $\alpha + \beta x$ and variance σ^2, the pdf for an observation is

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{[y_i - (\alpha + \beta x_i)]^2}{\sigma^2}\right) \qquad [2.7]$$

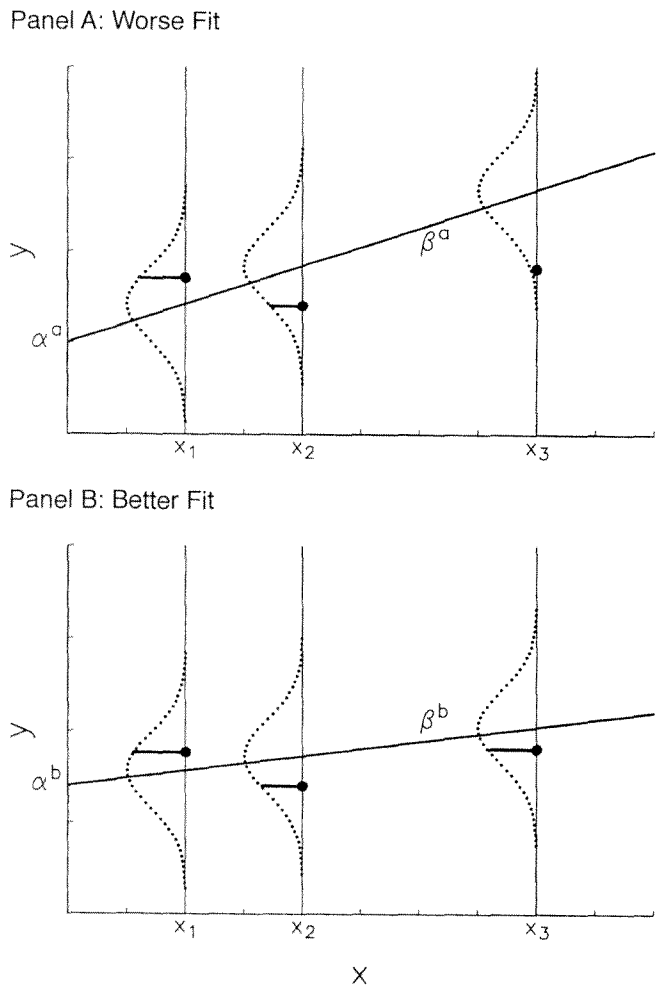


Figure 2.5. Maximum Likelihood Estimation for the Linear Regression Model

The pdf of a normal variable with mean μ and variance σ^2 is often expressed in terms of the pdf of a standardized normal variable ϕ with mean 0 and variance 1:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Using this definition, Equation 2.7 becomes

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)^2}{2}\right)\right]$$
$$= \frac{1}{\sigma}\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

and the likelihood equation can be written as

$$L(\alpha, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \frac{1}{\sigma}\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

Taking logs,

$$\ln L(\alpha, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma}\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right) \qquad [2.8]$$

ML estimates $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma}$ are obtained by maximizing this equation.

For multiple regression, $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon$ and

$$\ln L(\boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

The likelihood function is maximized when $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, which is the same as the OLS estimator. Maximum likelihood for the LRM is unusual since a *closed-form* solution is available. This means that the estimates can be obtained by algebraically solving the gradient of the log likelihood equation for the unknown parameters. Closed-form solutions are not possible for most of the models considered in later chapters and, consequently, iterative methods must be used. This topic is discussed in Chapter 3.

2.6.5. The Variance of ML Estimators

Maximum likelihood can also estimate the variance of the estimators. While the technical details are beyond the scope of our discussion (see Cramer, 1986, pp. 27–28; Davidson & MacKinnon, 1993, pp. 260–267; Eliason, 1993, pp. 40–41), we need a few definitions and results that are used in later chapters.

Let $\boldsymbol{\theta}$ be a vector containing the parameters being estimated. For example, in the simple regression $y = \alpha + \beta x + \varepsilon$ with $\text{Var}(\varepsilon \mid \mathbf{x}) = \sigma$, $\boldsymbol{\theta}$ will contain α, β, and σ. The *Hessian* is a matrix of second derivatives defined as

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

This is a square, symmetric matrix. For our example,

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\alpha} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\beta} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\sigma} \\ \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\alpha} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\beta} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\sigma} \\ \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\alpha} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\beta} & \dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\sigma} \end{pmatrix}$$

The second derivative indicates the rate at which the slope of the function is changing. For example, if $\partial^2 \ln L(\boldsymbol{\theta})/\partial\beta\partial\beta$ is small, then the log likelihood is changing slowly as β changes. That is, $\ln L$ is nearly flat. Intuitively, it makes sense that if $\ln L$ is flat, then it will be difficult to choose the value $\widehat{\beta}$ that maximizes the log likelihood. This should be reflected in the variance of $\widehat{\beta}$, since the variance reflects our certainty about the estimate. Indeed, the Hessian is related to the variance of the estimates through the information matrix.

The *information matrix* is defined as the negative of the expected value of the Hessian: $-E[\mathbf{H}(\boldsymbol{\theta})]$. Under very general conditions, the covariance matrix for the ML estimator is the inverse of the information matrix:

$$\text{Var}(\widehat{\boldsymbol{\theta}}) = -E[\mathbf{H}(\boldsymbol{\theta})]^{-1}$$

For our example,

$$\text{Var}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\alpha}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\beta}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\sigma}\right) \\ -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\alpha}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\beta}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\sigma}\right) \\ -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\alpha}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\beta}\right) & -E\left(\dfrac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\sigma}\right) \end{pmatrix}^{-1}$$

Various methods for estimating $\text{Var}(\widehat{\boldsymbol{\theta}})$ are considered in Chapter 3.

2.6.6. The Properties of ML Estimators

Under very general conditions, the ML estimator has a number of desirable properties. First, the ML estimator is *consistent*. This means roughly that as the sample size grows large, the probability that the ML estimator differs from the true parameter by an arbitrarily small amount tends toward 0. Second, the ML estimator is *asymptotically efficient*, which means that the variance of the ML estimator is the smallest possible among consistent estimators. Finally, the ML esimator is *asymptotically normally distributed*, which justifies the statistical tests that are discussed in Chapter 4. Notice that these are asymptotic properties, which means that they describe the ML estimator as the sample size approaches ∞. The degree to which they apply in finite samples is discussed in Section 3.5.

2.7. Conclusions

The linear regression model is our point of departure for presenting the models in later chapters. The next chapter begins by showing the problems inherent in using the LRM with a binary dependent variable. These problems lead to a latent regression model that generates the binary logit and probit models.

2.8. Bibliographic Notes

There are hundreds of texts dealing with the linear regression model. In order of increasing difficulty, I recommend Griffiths et al. (1993) for an introductory text; Kmenta (1986), Greene (1993), and Theil (1971) as intermediate texts; and Amemiya (1985) for an advanced treatment. Manski (1995) provides a detailed discussion of the identification problem. Four recommended sources on maximum likelihood, in order of increasing difficulty, are: Eliason (1993), Cramer (1986), Greene (1993, Chapter 12), and Davidson and MacKinnon (1993, Chapter 8).