Series/Number 07–146

LOGISTIC REGRESSION MODELS FOR ORDINAL RESPONSE VARIABLES

ANN A. O'CONNELL
University of Connecticut

Henry Gemurbon
september 2006

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

For information:

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

Sage Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

Sage Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

O'Connell, Ann A.
Logistic regression models for ordinal response variables / Ann
A. O'Connell.
p. cm. — (Quantitative applications in the social sciences; no. 146)
Includes bibliographical references and index.
ISBN 0-7619-2989-4 (pbk.)
1. Logistic regression analysis. 2. Social sciences—Statistical methods.
3. Educational statistics. I. Title. II. Series: Sage university papers
series. Quantitative applications in the social sciences; no. 146.
HA31.3.O27 2006
519.5′36—dc22

This book is printed on acid-free paper.

05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Lisa Cuevas Shaw
Editorial Assistant: Karen Gia Wong
Production Editor: Melanie Birdsall
Copy Editor: A. J. Sobczak
Typesetter: C&M Digitals (P) Ltd.
Indexer: Sheila Bodell

For Nathan, and especially for Delaney

CONTENTS

LIST OF TABLES AND FIGURES

Tables

Figures

SERIES EDITOR'S INTRODUCTION

Over the past three decades, logit type models have become the most popular statistical methods in the social sciences. In response to the need for understanding such models and showing how to correctly use them in various contexts, the Sage QASS (Quantitative Applications in the Social Sciences) series has given considerable attention to their exposition: The coverage includes No. 45 in the series, *Linear Probability, Logit, and Probit Models*, by Aldrich and Nelson; No. 86, *Logit Modeling*, by DeMaris; No. 101, *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*, by Liao; No. 106, *Applied Logistic Regression*, by Menard; No. 132, *Logistic Regression: A Primer*, by Pampel; No. 134, *Generalized Linear Models: A Unified Approach*, by Gill; No. 135, *Interaction Effects in Logistic Regression*, by Jaccard; and No. 138, *Logit and Probit: Ordered and Multinomial Models*, by Borooah. Why did my predecessor, Michael Lewis-Beck, who reviewed the prospectus and earlier drafts, put in the good work of editing another book on logit models for the series?

Since Rensis Likert's 1932 publication of *A Technique for the Measurement of Attitudes*, surveying human attitudes has never been the same. Indeed, any social surveys today will include the Likert-type scale as a staple means for asking questions. A typical Likert-type scale has five categories (e.g., *strongly disagree*, *disagree*, *undecided*, *agree*, *strongly agree*) to gauge one's response to a question, though it may have anywhere between three and seven or more response categories. If we code the five categories 1 to 5, we could estimate a linear regression model of a Likert-type scale, and that was the choice of method in the early days for analyzing such data. There are, however, some obvious problems. First and foremost, classical linear regression assumes a continuous dependent variable with equally spaced, ordered response categories. A Likert-type scale, or any other ordinal scale, is, albeit ordered, not necessarily equally spaced between categories. Second, and perhaps more important, such a scale would not give the normal distribution that the classical linear regression assumes the data to display.

To analyze ordinal data of this nature, there are other methods available, most often in the form of contingency tables and log-linear models. The Sage QASS series has also given attention to the topic, with the titles related to the topic including: No. 8, *Analysis of Ordinal Data*, by Hildebrand, Laing, and Rosenthal; No. 20, *Log-Linear Models*, by Knoke and Burke; No. 94, *Loglinear Models With Latent Variables*, by Hagenaars; No. 97, *Ordinal Log-Linear Models*, by Ishii-Kuntz; and No. 119, *Odds Ratios in the Analysis of Contingency Tables*, by Rudas. However, these methods are not in the regression framework, which is the most widely known and applied quantitative method in the social sciences.

Ann A. O'Connell's book fills the void. Even though Nos. 86, 101, and 138 in the series also treat ordered response variable in a logit model, the current book focuses entirely on such logit models by presenting three forms of the dependent variables that capture the ordinal nature of the response. The book begins by presenting an empirical example from the Early Childhood Longitudinal Study, for which the main dependent variable, although not a Likert scale, is nevertheless ordinal and measures proficiency in early literacy and numeracy. The author then reviews the logistic regression before presenting the core of the book in three topical chapters on the cumulative or proportional odds model, the continuation ratio model, and the adjacent categories model. Along the way, SAS® and SPSS® examples are given. Although the proportional odds model is perhaps the more widely applied of the three, the reader will appreciate the alternatives and especially the tips on when to use which, given in the concluding chapter.

—*Tim Futing Liao*
Series Editor

ACKNOWLEDGMENTS

Special thanks to Rosemarie L. Ataya for her initial and ongoing encouragement during the writing of this book, to D. Betsy McCoach for reading and rereading many drafts, and to several of my graduate students for their work on tables and some of the analyses: Heather Levitt Doucette, Jessica Goldstein, and Xing Liu. I would also like to thank John Fox, Scott Menard, and Timothy McDaniel for their reviews and valuable comments and suggestions, all of which greatly improved my own thinking and the quality of this work.

This research was supported in part by a grant from the American Educational Research Association, which receives funds for its "AERA Grants Program" from the National Science Foundation and the U.S. Department of Education's National Center for Education Statistics and the Office of Educational Research and Improvement (now the Institute for Education Sciences), under NSF Grant #REC-9980573. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

LOGISTIC REGRESSION MODELS FOR ORDINAL RESPONSE VARIABLES

Ann A. O'Connell
University of Connecticut

1. INTRODUCTION

For many response variables in education and the social sciences, ordinal scales provide a simple and convenient way to distinguish between possible outcomes that can best be considered as rank-ordered. The primary characteristic of ordinal data is that the numbers assigned to successive categories of the variable being measured represent differences in magnitude, or a "greater than" or "less than" quality (Stevens, 1946, 1951). Some examples of ordinal data include rubrics for scaling open-ended writing responses or essays and the solutions to arithmetic problems for which responses are scored based on improving levels of quality (e.g., 0 = poor, 1 = acceptable, 2 = excellent). In contrast, nominal-level data occur when the numeric values used to measure a variable simply identify distinct qualitative differences between categories (i.e., gender as 1 = male or 2 = female; geographic description of school attended as 1 = rural, 2 = urban, 3 = suburban, etc.); nominal data do not possess the directional characteristics of ordinal data. On the other hand, variables measured on an interval-level or ratio-level scale do use scale values to indicate the "greater than" or "less than" quality of ordinal variables but in addition maintain a property of equal-distance or equal-interval length between adjacent values across the scale. Temperature measured on the Celsius scale is a familiar example of an interval-level variable. However, interval-level variables have an arbitrary rather than an absolute zero-point. Variables that possess all the properties of interval scales but that also have a genuine zero-point are referred to as ratio-level; reaction time to a task, weight, and distance are familiar ratio-level variables.[1]

Ordinal categories are common in research situations where the assignment of numbers representing successive categories of an attribute, construct, or behavior coincides with meaningful directional differences. Knapp (1999) used ordinal ratings to assess severity of illness with scale categories such as mild (1), moderate (2), and severe (3). In Knapp's research, the numbers ascribed to the severity of illness categories represent increasing severity, in the sense that "moderate" is more critical than "mild," and "severe" is more critical than "moderate." The numerical rating given to the "severe" case does not imply that "severe" is three times as critical than "mild," only that the severity of illness in the "severe" category is greater than the severity of illness for those in the "mild" category, and greater still than those in the "moderate" category.

The choice of numbers used to represent the progressively more severe categories conveniently preserves the "greater than" or "less than" quality of the underlying attribute defining the categories themselves. The numbers model the attribute under study, such as severity of illness, and are chosen to preserve the transitivity of the categories: If the value of 3 represents a state that is more critical than the state represented by the value 2, and the value 2 represents a state more critical than the condition represented by the value 1, then the property of transitivity implies that the condition represented by the value of 3 is also more critical than the condition represented by the value of 1 (Cliff & Keats, 2003; Krantz, Luce, Suppes, & Tversky, 1971).

The measurement of variables on an ordinal scale is familiar. Ordinal scales have been used to categorize subjective probability or likelihood judgments in counseling and psychotherapy research (e.g., ratings from 1 = *very unlikely* to 5 = *very likely*) (Ness, 1995). A client's clinical condition after therapy can be characterized as deteriorated (1), unchanged (2), or improved (3) (Grissom, 1994). Health researchers frequently use five successive levels to characterize "stages of change" in health-related behavior such as smoking cessation, use of condoms, exercise behavior, and weight loss efforts (Hedeker & Mermelstein, 1998; Plotnikoff, Blanchard, Hotz, & Rhodes, 2001; Prochaska & DiClemente, 1983, 1986; Prochaska, DiClemente, & Norcross, 1992). In the stages-of-change model, disposition or activity toward behavior change typically is measured as precontemplation (1), contemplation (2), preparation (3), action (4), and maintenance (5). The experience of teachers' stages of concern for implementation of educational innovations in their classrooms has also been measured through an ordinal scale, one representing change in focus of concern from *self* = 1 to *other* = 7 (Hall & Hord, 1984; van den Berg, Sleegers, Geijsel, & Vandenberghe, 2000). In early-childhood education, indicators of mastery for the hierarchy of early literacy skills leading toward literacy proficiency in young children can be characterized as ordinal in nature: phonemic awareness (1), phonics (2), fluency (3), vocabulary (4), and text comprehension (5) (Center for the Improvement of Early Reading Achievement [CIERA], 2001).

Although ordinal outcomes can be simple and meaningful, their optimal statistical treatment remains challenging to many applied researchers (Cliff, 1996a; Clogg & Shihadeh, 1994; Ishii-Kuntz, 1994). Historically, researchers have relied on two very different approaches for the analysis of ordinal outcomes. Some researchers choose to apply parametric models for ordinal outcomes, such as through multiple linear regression with the outcome treated as at least an interval-level variable, assuming that the robustness of these techniques overcomes any potential interpretation problems. Other researchers choose to treat the ordinal variable as strictly categorical and apply log-linear or nonparametric approaches to understand the data. Although both strategies may be informative, depending on the research question, neither of these approaches is optimal for developing explanatory models of ordinal outcomes (Agresti, 1989; Cliff, 1996a; Clogg & Shihadeh, 1994; O'Connell, 2000), particularly when the focus of analysis is on the distinction between the ordinal scores.

Purpose of This Book

The purpose of this book is to familiarize applied researchers, particularly those within the fields of education and social and behavioral science, with alternatives for the analysis of ordinal response variables that are faithful to the actual level of measure of the outcome. The methods I discuss are examples of ordinal regression models, and they are extensions to logistic models for binary response data. Logistic regression methods are firmly established within epidemiology, medicine, and related fields, and in fact, much of the recent literature on application and development of ordinal regression techniques is found within the research of the larger public health community. Results of many of these statistical or comparative studies are mentioned here. Educational and social scientists may not typically focus on variables similar to those studied by epidemiologists or medical researchers, but both fields struggle with issues surrounding the aptness of models, and much can be learned about applications of different approaches to statistical dilemmas from the broader statistical literature.

In this book, three different methods for analyzing ordinal outcome data will be reviewed and illustrated through examples. These include the proportional or cumulative odds model (CO) (Agresti, 1996; Armstrong & Sloan, 1989; Long, 1997; McCullagh, 1980), the continuation ratio model (CR) (Armstrong & Sloan, 1989; D. R. Cox, 1972; Greenland, 1994), and the adjacent categories model (AC) (Agresti, 1989; Goodman, 1983). In addition, I present examples of partial proportional odds (Peterson & Harrell, 1990) and discuss the partial proportional hazards or unconstrained continuation ratio models (Bender & Benner, 2000; Cole & Ananth, 2001) as analysis alternatives for situations in which assumptions of the proportional odds or continuation ratio model are violated.

Ordinal logit models can be viewed as extensions of logistic regression for dichotomous outcomes, and consequently these models closely follow the approaches and model building strategies of both logistic and ordinary least squares regression analysis. I have chosen to focus on logit models for ordinal outcomes because the interpretations of probability and odds that derive from these models are somewhat intuitive. Alternatives to the methods presented here include, for example, Anderson's (1984) stereotype model, probit regression models, and the use of polychoric correlations for structural equation modeling of ordinal outcome variables. These and other strategies for analysis of ordinal data are discussed in Huynh (2002), Borooah (2002), Ishii-Kuntz (1994), Liao (1994), Menard (1995), and Jöreskog and Sörbom (1996); valuable references on the treatment of ordinal variables in general include Long (1997), Clogg and Shihadeh (1994), and Agresti (1989, 1996).

The cumulative odds model is the most frequently used ordinal regression model, although all of the models examined here are still relatively unfamiliar to many applied researchers, particularly in the educational sciences. Each of the models I review can address questions that are unique to the study of ordinal outcomes and that may not be satisfactorily answered by treating the data as either interval/ratio or strictly categorical.

Software and Syntax

The SAS® and SPSS® software packages are used for the examples presented here. Within each of these statistical packages, I used SAS PROC LOGISTIC (ascending and descending options), SAS PROC GENMOD, SAS PROC CATMOD, SPSS LOGISTIC REGRESSION, and SPSS PLUM to run the different models. Appendices in this book include the syntax used for each analysis presented, and both this syntax and the data can be found at my Web site (http://faculty.education.uconn.edu/epsy/aoconnell/index.htm). Limitations of, as well as similarities and differences between, the statistical packages will be noted as needed throughout this book. All analyses presented here assume independence across children. In the final chapter of this book, I briefly discuss the treatment of ordinal response variables for multilevel data, a rapidly building field that logically extends from work on the proportional odds model for single-level data and the fitting of multilevel models in general.

I focus on SAS and SPSS to illustrate the concepts and procedures for ordinal logit models included in this book. Another comprehensive

statistical package for the analysis of categorical data in general, one that contains excellent modules for analysis of ordinal data, is Stata (Long & Freese, 2003). Stata also includes graphical capabilities that can facilitate further understanding of the models presented here. The descriptions of the models included in this book are appropriate regardless of choice of statistical package.

Organization of the Chapters

Chapter 2 describes the data set used for the analyses presented here. Chapter 3 includes a brief review of logistic regression analysis, clarifying terminology important to the understanding of logit type ordinal regression models including odds, odd ratios, logits, and model fit. Each of the three ordinal models (CO, CR, AC) will then be described and illustrated in Chapters 4–6, building on their conceptual similarity to logistic regression models. For each of the ordinal models presented, model and variable effects will be explained, and assessment of model fit and predictive efficiency will be discussed. Chapter 4 provides a comparison with ordinary least squares multiple regression. Finally, Chapter 7 reviews and summarizes the analyses studied here and discusses some extensions to these models. Selected computer output will be included for each of the analyses presented.

The data for the examples contained in this book were drawn from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), which tracks the reading and arithmetic progress of a nationally representative sample of kindergarten children through the completion of first grade (third-grade data were released in March, 2004). Data from first-grade entry are analyzed here. The ECLS-K is conducted by the U.S. National Center for Education Statistics (NCES) and, in part, assesses student proficiency for early literacy, mathematics, and general knowledge as a series of "stepping-stones," which reflect the ordinal skills that form the foundation for further learning (West, Denton, & Germino-Hausken, 2000). All of the data are available on the first-grade public-use databases that can be obtained from NCES.[2] The examples illustrated here were derived solely for the purpose of explicating the technical and methodological use of ordinal regression models; although they are informative, they are not meant to provide a complete picture of early reading achievement for first-grade children. See, for example, Snow, Burns, and Griffin (1998) for further information about factors affecting early-childhood reading.

2. CONTEXT: EARLY CHILDHOOD LONGITUDINAL STUDY

Overview of the Early Childhood Longitudinal Study

The Early Childhood Longitudinal Study provides a comprehensive picture of first-grade children, their kindergarten and early home experiences, their teachers, and their schools. The ECLS-K investigates early literacy, reading, and arithmetic skills. It includes a battery of IRT (item-response theory)-scaled cognitive assessments collected on a nationally representative sample of approximately 20,000 children within sampled schools. In addition to the norm-referenced continuous IRT measures, the ECLS-K assesses criterion-referenced student proficiency for literacy and numeracy through responses to a series of five 4-item clusters that, as a set, reflect the skills that serve as stepping-stones for subsequent learning in reading and mathematics. The resulting scores can be used individually for student-level diagnosis and to identify directions for individualized interventions, as well as being used at a group level to suggest possible interventions for groups of students functioning at different levels of mastery. The analyses discussed in this book will focus on the criterion-referenced scores for literacy proficiency.

The categorization of early literacy proficiencies represented in the ECLS-K assessment instrument is consistent with the skills that have been identified as the building blocks of reading mastery: *phonemic awareness* (the understanding that letters represent spoken sounds), *phonics* (understanding the sounds of letters in combination), *fluency*, *vocabulary*, and *text comprehension* (CIERA, 2001). The skills underlying literacy development are hierarchical and interdependent; the later skills cannot realistically be expected to emerge without the development of the former. Table 2.1 describes the proficiency categories utilized by the ECLS-K.

The ability to respond sufficiently to the cluster of items represented by each category is assumed to follow the Guttman model (Guttman, 1954; NCES, 2000, 2002); that is, mastery at one level assumes mastery at all previous levels. On the ECLS-K assessments, a pass/fail score was obtained for each child in the sample on each cluster of items representing a proficiency level (1 through 5) until the child failed to pass three out of the four items in a cluster.[3] Mastery of one cluster indicates mastery of all previous clusters; testing was stopped once a child was unable to successfully pass a cluster of items.[4] Consequently, there are five dichotomous variables for literacy proficiency (C3RRPRF1 to C3RRPRF5) in the ECLS-K database. For example, if a child passes three out of four items in literacy level 1 and

TABLE 2.1
Proficiency Categories for the ECLS-K Measures for Early Literacy

Proficiency Category	*Description*
0	Did not pass level 1
1	Can identify upper/lowercase letters
2	Can associate letters with sounds at the beginnings of words
3	Can associate letters with sounds at the ends of words
4	Can recognize sight words
5	Can read words in context

SOURCE: National Center for Education Statistics (2002).

three out of four items in literacy level 2, that child would receive a value of 1 for both C3RRPRF1 and C3RRPRF2. If this same child does not pass three out of four items in the next cluster (literacy level 3), a score of 0 is recorded for C3RRPRF3 as well as for all subsequent levels. For the analyses presented here, the series of five dichotomous proficiency values was used to create a single variable that reflects mastery of the content areas on an ordinal scale. After recoding to achieve a single ordinal variable, the hypothetical student above would receive a value of 2 as his or her proficiency score, representing mastery of material up to and including level 2. In this manner, a single variable (*profread*) with six possible outcome categories (levels 0 through 5) for the assessment of literacy proficiency was derived for each child in the ECLS-K sample. A score of 0 on this ordinal scale implies that the child did not attain mastery for the cluster of items representing proficiency level 1.[5]

Practical Relevance of Ordinal Outcomes

Ordinal proficiency scores can reveal to researchers and educators how far along children are on the path to becoming fully literate as they continue through their primary school education. Analyzing the ordinal proficiency scores rather than the continuous IRT-scaled scores as the variables of interest highlights the role that proficiency assessments can play in the identification and selection of students for early intervention programs. These analyses can suggest concrete areas in the hierarchy where interventions might be tailored to meet particular student needs. Ordinal proficiency outcomes, and indeed ordinal variables in general, have a great deal of pragmatic utility in the degree to which they can direct intervention to specific levels of proficiency. For the classroom teacher or reading specialist, proficiency scores may be far more valuable, and interpretable, than knowing that a child's IRT-scaled score on a cognitive assessment is "55." Interventions tailored to the classroom, or school practices or policies found to be associated with the stepping-stones to successful acquisition of literacy skills, may be far more effective for individual students than strategies based on attempts to improve a global cognitive test score (obtained at the classroom, school, or district level).

Variables in the Models

The variables selected as predictors in the analyses presented here have been found to be associated with early reading skill among young children. Initial data summaries of the ECLS kindergarten cohort indicate that some children do enter kindergarten with greater preparedness and "readiness" to learn than that exhibited by other children, perhaps putting them a step ahead of their peers for the important early grades at school (NCES, 2000). ECLS-K studies have shown that children entering kindergarten who have particular characteristics (living in a single-parent household, living in a family that receives welfare payments or food stamps, having a mother with less than a high school education, or having parents whose primary language is not English) tended to be at risk for low reading skills (Zill & West, 2001). Pre-kindergarten experiences related to family life (e.g., being read to by parents), attendance at preschool or day care, and personal characteristics (e.g., gender) may relate to children's initial proficiency in reading as well as to their potential growth in skills and abilities across the kindergarten year and beyond. For example, girls typically enter kindergarten with slightly greater early literacy ability than boys. Child-focused predictors of success and failure in early reading are helpful for understanding how individual children may be at risk for reading difficulties. From a policy and practice perspective, it is clearly desirable that teachers, school administrators, parents, and other stakeholders be aware of these individual factors related to early proficiency so that these stakeholders can develop and support curriculum and instructional practices that can promote achievement for all students relative to their first-grade and kindergarten entry skills.

Descriptive statistics for the explanatory variables across the six proficiency categories are presented in Table 2.2. These include *gender*, shown here as *% male* (0 = female, 1 = male), *risknum* (number of family risk characteristics, ranging from 0 to 4, based on parent characteristics including living in a single-parent household, living in a family that receives welfare payments or food stamps, having a mother with less than a high school education, or having parents whose primary language is not English), *famrisk* (dichotomous variable indicating whether or not any family risk was present, coded 0 = no, 1 = yes [or *risknum* greater than or equal to 1]), *p1readbo* (frequency with which parents read books to children prior to kindergarten entry, rated as 1 to 4 with 1 = never and 4 = every day), *noreadbo* (dichotomized variable indicating 0 = parent reads books to child three or more times a week to every day and 1 = parent reads books to child less than once or twice per week), *halfdayK* (child attended half-day versus full-day kindergarten, coded 0 = no [attended full-day K], 1 = yes [attended half-day K]), *center* (whether or not child ever received center-based day care prior to attending kindergarten; 0 = no, 1 = yes), *minority* (0 = white/Caucasian background; 1 = minority [any other] background), *wksesl* (family SES assessed prior to kindergarten entry, continuous scaled score with mean of 0), and *p1ageent* (age of child in months at kindergarten entry). An additional variable, included for descriptive purposes but not included in the models because of design concerns, is *public* (type of school child attended, rated as 0 = private, 1 = public).

TABLE 2.2
Descriptive Statistics at First-Grade Entry, $N = 3,365$

	Reading Proficiency Level (profread)						
	0 (*n = 67*)	*1* (*n = 278*)	*2* (*n = 594*)	*3* (*n = 1,482*)	*4* (*n = 587*)	*5* (*n = 357*)	*Total* (*N = 3,365*)
% profread	2.0%	8.3%	17.7%	44.0%	17.4%	10.6%	100%
% male	71.6%	58.6%	53.9%	49.6%	43.6%	42.3%	49.7%
risknum							
M	.97	.77	.65	.44	.32	.25	.47
(SD)	(1.04)	(0.88)	(0.88)	(0.71)	(0.61)	(0.53)	(0.75)
% famrisk	58.2%	52.9%	43.8%	32.5%	25.9%	20.7%	34.3%
% noreadbo	38.8%	27.0%	21.7%	15.5%	13.1%	7.6%	16.7%
% halfdayK	43.3%	41.7%	46.3%	48.0%	40.7%	43.7%	45.3%
% center	71.6%	73.7%	71.0%	77.5%	78.7%	84.9%	76.9%
% minority	59.7%	58.3%	48.5%	33.3%	34.2%	33.9%	38.8%
wksesl							
M	−.6133	−.2705	−.1234	.1490	.2807	.6148	.1235
(SD)	(0.67)	(0.64)	(0.71)	(0.75)	(0.70)	(0.75)	(0.76)
p1ageent							
M	65.6	65.1	65.5	66.1	66.5	67.1	66.1
(SD)	(4.40)	(4.34)	(3.97)	(4.00)	(4.07)	(3.86)	(4.06)
% public	98.5%	93.5%	86.9%	76.7%	70.9%	61.9%	77.7%

The design of the ECLS-K sampling plan called for oversampling of children with Asian and Pacific Islander backgrounds, and it currently includes three waves of data, collected at kindergarten entry, at the end of the kindergarten year, and at the end of the first-grade year. Third-grade data were released in the spring of 2004. Data also were collected on a 30% subsample of children at first-grade entry. All data used for the examples presented here were contained in the 30% first-grade subsample; the children had no missing data on the variables of interest, were first-time kindergarteners (no repeaters), and remained in the same school for first grade that they attended in kindergarten. Given the focus of this book and the oversampling of Asian/Pacific Islanders, coupled with sparse cells for other minority groups, a dichotomous variable for race/ethnicity was created with a classification of 1 = minority group and 0 = white/Caucasian for these illustrative models. With this criteria, there were n = 3,365 children from 255 schools (57 private and 198 public), with an average of 13 students per school. Incorporating the nested design into the analysis of ordinal outcome data is addressed in Chapter 7; all other analyses assume independence of children across schools.

3. BACKGROUND: LOGISTIC REGRESSION

Overview of Logistic Regression

Ordinal regression models are closely related to logistic models for dichotomous outcomes, so I begin with a brief review of logistic regression analysis in order to highlight similarities and differences in later chapters. Other authors in the QASS series and elsewhere (e.g., Cizek & Fitzgerald, 1999; Hosmer & Lemeshow, 1989, 2000; Menard, 1995, 2000; Pampel, 2000) have covered logistic regression in depth, so only those concepts important to the discussion later in this book are included here.

The terminology and estimation strategies for fitting ordinal regression models are fairly straightforward extensions of those used for logistic regression. These models are collectively defined as a class of generalized linear models, consisting of three components:

- A random component, where the dependent variable Y follows one of the distributions from the exponential family such as the normal, binomial, or inverse Gaussian

- A linear component, which describes how a function, Y', of the dependent variable Y depends on a collection of predictors
- A link function, which describes the transformation of the dependent variable Y to Y' (Fox, 1997).

The identity link function does not alter the dependent variable, leading to the general linear model for continuous outcomes, for which multiple linear regression is the familiar case. The logit link function transforms the outcome variable to the natural log of the odds (explained below), which leads to the logistic regression model.

Logistic analyses for binary outcomes attempt to model the odds of an event's occurrence and to estimate the effects of independent variables on these odds. The odds for an event is a quotient that conveniently compares the probability that an event occurs (referred to as "success") to the probability that it does not occur (referred to as "failure," or the complement of success). When the probability of success is greater than the probability of failure, the odds are greater than 1.0; if the two outcomes are equally likely, the odds are 1.0; and if the probability of success is less than the probability of failure, the odds are less than 1.0.

For the ECLS-K example described above, suppose we are interested in studying the attainment of reading proficiency category 5 (sight words) among children at first-grade entry. The outcome can be described as binary: A child attains proficiency in category 5 (success) or not (failure). The odds of reaching category 5 are computed from the sample data by dividing the probability of reaching category 5 (scored as Y = 1) by the probability of not reaching category 5 (scored as Y = 0):

$$\text{Odds} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}.$$

To examine the impact on the odds of an independent variable, such as gender or age, we construct the odds ratio (OR), which compares the odds for different values of the explanatory variable. For example, if we want to compare the odds of reaching proficiency category 5 between males (coded x = 1) and females (coded x = 0), we would compute the following ratio:

$$OR = \frac{\dfrac{P(Y=1|x=1)}{1-P(Y=1|x=1)}}{\dfrac{P(Y=1|x=0)}{1-P(Y=1|x=0)}}.$$

Odds ratios are bounded below by 0 but have no upper bound; that is, they can range from 0 to infinity. An OR of 1.0 indicates that an explanatory variable has no effect on the odds of success; that is, the odds of success for males is the same as the odds of success for females. Small values of the OR (< 1.0) indicate that the odds of success for the persons with the value of x used in the denominator (0 = females) are greater than the odds of success for the persons with the higher value of x used in the numerator (1 = males). The opposite is true for values of the OR that exceed 1.0; that is that the odds for males of being in proficiency category 5 is greater than the odds for females. The nature and type of coding used for the independent variables become important in interpretation; in this example and throughout this text, I used simple dummy or referent coding. Other approaches to coding categorical independent variables can change the interpretation of that variable's effect in the model; discussions of alternative approaches to categorizing qualitative data in logistic regression models can be found in Hosmer and Lemeshow (2000).

The OR is a measure of association between the binary outcome and an independent variable that provides "a clear indication of how the risk of the outcome being present changes with the variable in question" (Hosmer & Lemeshow, 1989, p. 57). Although the probability of an event could be modeled directly through the linear probability model (i.e., using ordinary linear regression on the dichotomous [0, 1] dependent variable), such an approach leads to some serious interpretation problems. The linear probability model can yield implausible predictions outside the 0, 1 bounds for probability, particularly if the independent variable is continuous. In addition, the typical assumptions of homoscedasticity and normality of errors from the ordinary linear regression model are violated when the outcome is dichotomous, calling the validity of results from such an approach into question (Cizek & Fitzgerald, 1999; Ishii-Kuntz, 1994; O'Connell, 2000). Instead, when the outcome is dichotomous, we model the odds, or more specifically, we model the natural (base e) log of the odds, referred to as the *logit* of a distribution.

This simple transformation of the odds has many desirable properties. First, it eliminates the skewness inherent in estimates of the OR (Agresti, 1996), which can range from 0 to infinity, with a value of 1.0 indicating the *null* case of *no change in the odds*. The logit ranges from negative infinity to infinity, which eliminates the boundary problems of both the OR and probability. The transformed model is linear in the parameters, which means that the effects of explanatory variables on the log of the odds are additive. Thus, the model is easy to work with and allows for interpretation of variable effects that are exceptionally straightforward, and for model-building strategies that mirror those of ordinary linear regression.

This process can be extended to include more than one independent variable. If we let $\pi(Y = 1|X_1, X_2, \ldots X_p) = \pi(\underline{x})$ represent the probability of "success," or the outcome of interest (e.g., a child being in proficiency category 5), for a given set of p independent variables, then the logistic model can be written as

$$\ln(Y') = \text{logit}\,[\pi(\underline{x})] = \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right)$$
$$= \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

In this expression, Y' is simply a convenient way to refer to the odds in the transformed outcome variable; rather than predicting Y directly, we are predicting the (log of the) odds of $Y = 1$. The link function describes the process of "linking" the original Y to the transformed outcome: $f(y) = \ln(Y')$ $= \ln[\pi(\underline{x})/(1 - \pi(\underline{x}))]$, which is referred to as the logit link. Solving for $\pi(\underline{x})$ gives us the familiar expression for the logistic regression model for the probability of success:

$$\pi(\underline{x}) = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}$$
$$= \frac{1}{1 + \exp[-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)]}.$$

Statistical packages such as SPSS and SAS provide maximum likelihood (ML) estimates of the intercept and regression weights for the variables in the model. Maximum likelihood estimates are derived using an iterative method that returns the "values for the population parameters that 'best' explain the observed data" (Johnson & Wichern, 1998, p. 178). These ML estimates maximize the likelihood of obtaining the original data, and because the logistic model is developed through a nonlinear transformation of the outcome, the method does not require a normal distribution of the error terms, as does ordinary least squares estimation. The likelihood represents the probability that the observed outcomes can be predicted from the set of independent variables. Likelihood can vary between 0 and 1; the log-likelihood (LL) varies from negative infinity to 0. Multiplying the LL by −2 creates a quantity that can be used for hypothesis testing purposes to compare different models (Hosmer & Lemeshow, 2000).

Assessing Model Fit

One way to assess how well a fitted model reproduces the observed data is to compute the "deviance" for the fitted model. The deviance represents how *poorly* the model reproduces the observed data, and it is found by comparing the likelihood of the fitted model to a model that has a perfect fit, called the saturated model.[6] The saturated model has as many parameters as there are values of the independent variable; the likelihood of the saturated model is 1.0, and –2LL(saturated model) = 0. The "deviance" of any model, D_m, is thus the quantity –2LL (see Hosmer & Lemeshow, 2000). We would expect the "poorness" of fit to decrease (toward 0) with better-fitting models. The fit of two nested models, with variables in Model 1 a subset of those in Model 2, can be compared by considering the difference of their deviances: $G = D_{m1} - D_{m2}$. The quantity G represents "goodness" of fit, and for large samples, G follows an approximate chi-square distribution with degrees of freedom equal to the difference in number of parameters estimated between Model 1 and Model 2. A statistically significant G indicates that Model 2 has less "poorness" of fit than Model 1.

When Model 1 is the null model, this comparison provides an omnibus test (assuming large-sample properties and non-sparse cells) for whether or not the fitted model reproduces the observed data better than the null, or intercept only, model. However, it does not tell us how well the model performs relative to the saturated, or perfect, model. With categorical predictors, SAS tests D_m (which compares the fitted to the saturated model) using the Pearson χ^2 criteria or the Deviance χ^2 criteria. Neither of these is appropriate when continuous explanatory variables are included (see Allison, 1999; Hosmer & Lemeshow, 2000). When explanatory variables are categorical, these tests can be generated in SAS using the "*/aggregate scale=none*" option in the model statement.

With small samples or when sparse cells are present in the data (which nearly always will occur with the inclusion of continuous independent variables in the model), alternative methods for assessing model fit should be considered; a common strategy is known as the Hosmer-Lemeshow (H-L) test (1989, 2000). The H-L test is obtained through SAS by requesting the "*/lackfit*" option in the model statement; in SPSS, the test is provided when "*goodfit*" is included in the print statement.

The H-L test works well when independent variables (IVs) are continuous, because it deals directly with the number of covariate patterns within the data. When IVs are continuous, there is essentially a different possible covariate pattern for each observation in the data set. Briefly, the H-L test forms several groups referred to as "deciles of risk" based on the estimated probabilities for the sample. In most situations, g = 10 groups are formed, but there may be fewer depending on similarity of estimated probabilities across different covariate patterns. The cases within these deciles are then used to create a g × 2 table of observed to expected frequencies, and a Pearson χ^2 statistic is calculated for this table (Hosmer & Lemeshow, 1989, 2000). If the model fits well, agreement is expected between the observed and expected frequencies, so that the null hypothesis of a good fit between observed and expected frequencies from the model would be retained. The H-L test has been criticized in the literature for lack of power (Allison, 1999; Demaris, 1992), but reliance on a single test to indicate model adequacy is in itself discouraged (Hosmer & Lemeshow, 2000).[7] Supplemental strategies include measures of association and predicative efficiency, discussed later in this chapter.

Interpreting the Model

Typically, SPSS models the log of the odds for the dependent variable coded with the higher value (the 1, if the outcome is coded as 0 or 1), but SAS by default models the response coded with the lower value. With binary outcomes, the interpretation of results and effects of independent variables on the odds is not affected by decisions of how "success" versus "failure" are coded, because these two events are complements of each other. For example, let the probability of "success" as defined by P(reaching proficiency category 5) = .2. Then, the probability of "failure" or P(not reaching proficiency category 5) = 1 – .2 = .8. The odds of success would then be .25 (.2/.8). The odds for the complement of the event, which is not reaching proficiency category 5, would be 1/.25 or 4.0 (.8/.2). Because there are only two possible outcomes for the dependent variable, the odds for the complement of an event is simply the inverse of the odds for that event. When the logistic transformation is applied, we see that taking the log of the odds of an event (ln(.25) = –1.3863) has the opposite sign, but the same magnitude, of the log of the odds for the complement of the event (ln(4) = +1.3863). In the logistic regression model, reversing the coding for the outcome being modeled amounts to the same probability predictions and interpretations once the direction of the regression coefficients and the intercept are taken into account. With dichotomous outcomes, use of the "descending" option in the model statement for SAS changes the default approach and asks the computer to model the odds for the higher-valued outcome category, which would be the category labeled $Y = 1$ if the outcomes are coded as 0 or 1 (or category 2 if the outcomes are labeled as 1 and 2). However, with more than two *ordinal* response categories, applying the "descending" option can change the model dramatically and must be used with care. Use of this option for ordinal outcomes will be explained fully in Chapter 4.

For $j = 1$ to p independent variables, the regression weights in the multivariate logistic model represent the *change in the logit* for each one-unit increase in X_j, controlling or adjusting for the effects of the other independent variables in the model. Because it is more intuitive to consider variable effects in terms of the odds rather than the log-odds (the regression weights are in log-odds), information about the odds themselves is found by exponentiating the weights for the variables in the model (i.e., exp(b*j*)). The exponentiations of the regression weights are the ORs and are routinely reported in computer runs. The ORs can be interpreted directly to indicate the effect of an independent variable on the odds of success, and the percentage change in the odds also can be calculated using the following formula: (100 × [OR – 1]).

Strong associations between independent variables and the outcome typically are represented by ORs farther from 1.0, in either direction. Long (1997) refers to the ORs as "factor change" estimates (p. 79). For a unit change in the independent variable, the corresponding OR is the factor by which the odds of "success" are expected to change, controlling for all other independent variables in the model. Statistical significance of an OR typically is assessed by testing if the regression coefficient, β_j, is statistically different from zero through one of three approaches: a Wald, score, or likelihood ratio test. In the Wald test, the parameter estimate for the effect of each independent variable in a logistic model is divided by its respective standard error, and the results are squared to represent a value from the chi-square distribution with one degree of freedom under the null hypothesis of no effect. However, the Wald statistics can be problematic in small samples; in samples with many different data patterns, such as when an independent variable is continuous rather than categorical; or in samples with sparse cells for categorical IVs (Jennings, 1986; Menard, 1995). Both SPSS and SAS report Wald chi-square statistics for each variable in the fitted model.

The score test for the contribution of an independent variable in the model relies on derivatives of the likelihood function and is not directly available in either SPSS or SAS; however, SPSS does use a score test in stepwise procedures to determine when variables enter or exit a developing model (Hosmer & Lemeshow, 2000). The likelihood ratio test has been advocated as the most reliable test for contribution of an independent variable to a model, but it is not directly available in either SPSS or SAS. The test can be obtained easily through some simple but possibly time-consuming programming, and it involves comparing the deviances for nested models, that is, the deviance from a model that does not contain the independent variable of interest to the deviance of a model that does. The difference in deviances approximates a chi-square distribution with one degree of freedom. Because the focus of this book is on development and overall interpretation of ordinal models, I chose to rely on the Wald test for assessing effects of explanatory variables. However, researchers do need to be aware that alternatives to this test exist.

Measures of Association

There are several logistic regression analogs to the familiar model R^2 from ordinary least squares regression that may be useful for informing about strength of association between the collection of independent variables and the outcome, although Menard (2000) and others (Borooah, 2002; Demaris, 1992; Long, 1997) point out that there is some disagreement among researchers as to which proportion reduction in error measure is most meaningful. For logit type models, the likelihood ratio R^2 value, R_L^2, seems to provide the most intuitive measure of improvement of fit for a multivariate model relative to the null (intercept only) model. R_L^2 is found by comparing two log-likelihoods: $R_L^2 = 1 -$ (log-likelihood(model)/log-likelihood(null)) (Hosmer & Lemeshow, 2000; Long, 1997; McFadden, 1973; Menard, 2000). It measures the proportion reduction of error (log-likelihood) achieved from the use of the set of independent variables (relative to the null model). Other alternatives for measuring strength of association exist, but only a few will be discussed in the examples to follow. Long (1997) states that "While measures of fit provide some information, it is only partial information that must be assessed within the context of the theory motivating the analysis, past research, and the estimated parameters of the model being considered" (p. 102). The interested reader should consult Menard's (2000) discussion on the use of various R^2 analogs in logistic regression, as well as Borooah (2002, pp. 19–23). Huynh (2002) provides a discussion of extensions of these situations in which the outcome is ordinal rather than dichotomous.

Example 3.1: Logistic Regression

A simple example will be used to illustrate the concepts above, as well as to provide an extension for developing an ordinal regression model. I chose a subset of the original ECLS-K data described above: $n = 702$ children who fell into proficiency categories 0, 1, or 5 when they were tested at the beginning of first grade. Table 3.1 provides the frequency breakdown for this subsample according to gender. The subsample is fairly balanced

TABLE 3.1
Cross-Tabulation of Proficiency (0, 1 versus 5) by Gender, $N = 702$

Gender	Y = 0 (profread category 0 or 1)	Y = 1 (profread category 5)	*Totals*
Males ($x = 1$)	211	151	362
Females ($x = 0$)	134	206	340
Totals	345	357	702

across the two outcomes. In the data analysis to follow, males were coded as "$x = 1$" and females as "$x = 0$," with the outcome of being in category 5 coded as "$Y = 1$" and being in either category 0 or 1 coded as "$Y = 0$."

The odds for a male being in the higher proficiency category can be found by dividing the probability of being in category 5 by the probability of not being in category 5:

$$\text{Odds (category 5|male)} = \frac{151/362}{211/362} = \frac{.4171}{1 - .4171} = .7156.$$

Similarly for females, the odds of being in proficiency category 5 are determined as

$$\text{Odds (category 5|female)} = \frac{206/340}{134/340} = \frac{.6059}{1 - .6059} = 1.537.$$

From these two values, we see that for this subsample, boys have a greater probability of being in categories 0 or 1 rather than in category 5 (the numerator is less than .5), and for girls, the opposite is true (the numerator is greater than .5). Thus, the odds for a boy of being in category 5 is less than the odds for a girl of being in category 5. The odds ratio (OR) compares these two odds and provides a measure of the association between gender and the odds of being in category 5:

$$\text{OR} = \frac{\text{Odds (category 5|male)}}{\text{Odds (category 5|female)}} = \frac{.7156}{1.537} = .466.$$

The OR of .466 informs us that, for this subsample, the odds for boys being in the higher proficiency category is .466 times the odds for girls of being in category 5, or less than half. Put another way, being a boy decreases the odds of being in category 5 by 53.4% (100 × [OR − 1] = −53.4). Conversely, the odds for a girl of being in category 5 is 2.146 times the odds for boys, or more than twice the odds for boys (1/.466 = 2.146).

In a logistic regression model, as discussed earlier, probability is transformed to the odds, and the odds are transformed to logits by taking the natural log. Selected output from fitting the logistic regression model for the above example using SPSS LOGISTIC REGRESSION is shown in Figure 3.1 (syntax in Appendix A, section A1). In this model, Y is coded 1 for being in proficiency category 5, and 0 if not. The explanatory variable, "*gender*," is coded 1 if the child is a boy, and 0 if the child is a girl. We will let $\ln(Y')$ represent the logit, or log-odds. The prediction model is $\ln(\hat{Y}') = .430 + (-.765)$ *gender*. Parameter estimates are found in the last section of Figure 3.1, "Variables in the Equation."

When the child is female (*gender* = 0), the constant represents the prediction for the log of the odds; it is .430. Exponentiating this back to the odds, we have exp(.430) = 1.537, which is, as solved for above, the odds of being in proficiency category 5 for a girl. For boys (coded *gender* = 1), our model's prediction becomes .430 + (−.765 × 1) = −.335. Exponentiating this result, we have exp(−.335) = .7153, which is (within rounding error) the odds of being in proficiency category 5 for a boy. Finally, the OR (taking rounding into consideration) can be found by exponentiating the regression weight for *gender*, exp(−.765) = .466. This value appears in the final column of the "Variables in the Equation" table, and it is precisely the OR determined from the frequency data. It tells us that the odds of being in proficiency category 5 for a boy is .466 times the odds for a girl.

For many researchers, it is easier to interpret the OR than to interpret the logits, but the logits can also be interpreted directly. The effect for *gender* in the logistic regression model tells us how much the logit is expected to change when the value for *gender* changes by one unit, in this case from 0 (female) to 1 (male). Based on the Wald criteria, the effect of *gender* is statistically significant in the logit model: Wald's $\chi^2_1 = 24.690$, $p = .000$. This implies that the estimated slope for *gender* is −.765 and is statistically different from 0, and that the OR = exp(−.765) = .466 is therefore statistically different from 1.0.

In this SPSS example, the deviance of the null model is found in the section for Block 1, "Iteration History," footnote c of Figure 3.1: $D_0 = -2LL_0 = 972.974$. The deviance of the fitted model containing only the variable *gender* is $D_m = -2LL_m = 947.825$. The difference between these two deviances is $G_m = 25.149$, $df = 1$, $p = .000$. For this example, with only one independent variable included in the model, the omnibus test is also the likelihood ratio test (an alternative to the Wald χ^2 test) for the effect of *gender*. The

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	702	100.0
	Missing Cases	0	.0
	Total	702	100.0
Unselected Cases		0	.0
Total		702	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		−2 Log-likelihood	Coefficients Constant	gender
Step 1	1	947.829	.424	−.755
	2	947.825	.430	−.765
	3	947.825	.430	−.765

a. Method: Enter
b. Constant is included in the model.
c. Initial −2 Log-Likelihood: 972.974
d. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-Square	df	Sig.
Step 1	Step	25.149	1	.000
	Block	25.149	1	.000
	Model	25.149	1	.000

Figure 3.1 Selected Output: SPSS Logistic Regression Example

Figure 3.1 (Continued)

Model Summary

Step	−2 Log-likelihood	Cox & Snell *R* Square	Nagelkerke *R* Square
1	947.825	.035	.047

Classification Table[a]

Observed			Predicted CUMSP2 .00	CUMSP2 1.00	Percentage Correct
Step 1	CUMSP2	.00	211	134	61.2
		1.00	151	206	57.7
	Overall Percentage				59.4

a. The cut value is .500.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B) Lower	Upper
Step 1[a]	gender	−.765	.154	24.690	1	.000	.466	.344	.629
	Constant	.430	.111	15.014	1	.000	1.537		

a. Variable(s) entered on step 1: gender.

omnibus test, found in "Omnibus Tests of Model Coefficients," means that we find a statistically significant decrease in the −2LL when *gender* is included in the model. This reduction represents a proportionate reduction in deviance that can be expressed through the likelihood ratio R^2_L: $1 - (D_m/D_0) = .0258$. For this model, the inclusion of *gender* in the model reduces the deviance of the null model ($D_0 = -2LL_0$) by 2.58%.

Neither SPSS nor SAS reports R^2_L in their logistic regression procedures, but as shown above, it can be calculated easily from the available statistics provided in either package. Both statistical packages report two variations on the R^2 statistic for logit analysis: the Cox and Snell R^2, which SAS reports as the (generalized) R^2, and the Nagelkerke R^2, which SAS refers to as the "max-rescaled R^2." The Nagelkerke R^2 rescales the Cox and Snell R^2 value to obtain a bound of 1.0. For these data, the "Model Summary" table of Figure 3.1 reports $R^2_{CS} = .035$ and $R^2_N = .047$. Although the omnibus test is statistically significant, none of the R^2 statistics is very large, suggesting that other explanatory variables in addition to *gender* may be helpful in understanding the likelihood of a child being in proficiency category 5. Menard (2000) discusses several attempts to generalize the familiar R^2 from ordinary linear regression, but he advocates R^2_L as the most useful of the available pseudo R^2's.

Attempting to reduce the fit assessment to a single value, as the collection of pseudo R^2's do, may have value in terms of comparing across competing (nested) models, but this provides only a "rough index of whether a model is adequate" (Long, 1997, p. 102). An investigation of model adequacy can be augmented by assessing how well the observed categorical outcomes are reproduced, based on whether or not an individual is predicted to fall into his or her original outcome of $Y = 0$ or $Y = 1$. This assessment of predictive efficiency supplements the information available from the tests for model fit and the reduction in deviance statistics. Some measures of fit or correspondence between observed and predicted outcomes are strongly influenced by data that are highly unbalanced in terms of distribution of frequency of the outcome, so an informed decision is best made by computing and comparing across several different measures rather than relying on one single measure.

To consider the ability of a model to correctly classify cases, classification is based on the probabilities estimated from the model, and the results are compared with the observed frequencies for each category. For any child, if the probability of "success" based on the logistic model is greater than .5, the predicted outcome would be 1; or else the predicted outcome would be 0 (Hosmer & Lemeshow, 2000; Long, 1997). SPSS produces a classification table directly, shown under Block 1: "Classification Table." The predicted probabilities can be requested in SAS (as well as in SPSS) to construct the classification table; review the syntax in Appendix A, sections A1 and A2, for how to save these predicted probabilities. Although many different kinds of classification statistics are available (Allison, 1999; Gibbons, 1993; Hosmer & Lemeshow, 2000; Huynh, 2002; Liebetrau, 1983; Long, 1997; Menard, 1995, 2000), several seem to be reported in the literature in preference to others and can be used with ordinal dependent variables. These include τ_p, which "adjusts the expected number of errors for the base rate of the classification" (Menard, 1995, p. 29), and the *adjusted count* R^2 or $R^2_{adjCount}$, which is similar to the Goodman-Kruskal λ in its asymmetric form (that is, when one variable is being predicted from a set of other variables); $R^2_{adjCount}$ adjusts the raw percentage correct measure for the likely probability of a case being assigned to the modal category of

the observed dependent variable (DV) by chance (Liebetrau, 1983; Long, 1997). Unfortunately, there are often several different names for the same measures within the literature, and the reader of multiple articles or texts should pay close attention to the nomenclature that each author uses. For example, Menard (1995, 2000) refers to $R^2_{adjCount}$ as λ_p.

Hosmer and Lemeshow (2000) point out that model fit in terms of correspondence between observed and estimated probabilities is often more reliable and meaningful than an assessment of fit based on classification. They suggest that classification statistics be used as an adjunct to other measures, rather than as a sole indicator of quality of the model. As mentioned above, multiple criteria for investigating adequacy of fit of the models are demonstrated and reported in the examples covered here.

Neither SAS nor SPSS provides τ_p or λ_p ($R^2_{adjCount}$) directly, but they can be calculated once the classification table is obtained. To find τ_p, the expected number of errors must first be determined, and for 2 × 2 tables, this is

$$E(\text{errors}) = 2 \times \frac{f(Y=0) \times f(Y=1)}{n}.$$

The desired measure of association can then be calculated from

$$\tau_p = \frac{E(\text{errors}) \times O(\text{errors})}{E(\text{errors})}.$$

The observed errors are the off-diagonal elements of the classification table. A different expression for τ_p can be found in Menard (2000); it is also appropriate for ordinal response models:

$$\tau_p = 1 - \frac{\left(n - \sum_i f_{ii}\right)}{\sum_i \frac{f_i(n - f_i)}{n}},$$

where i represents the index for each category of the outcome variable, n = sample size, f_{ii} = sum of the correctly predicted categories (on the diagonal of the classification table), and f_i = the observed frequency for category i. For these data, $\tau_p = .1878$, indicating that after adjustment for the base rate, classification error is reduced by approximately 19% using the model with *gender* as the only predictor.

To find the $R^2_{adjCount}$ or λ_p for the classification table, following Long (1997) and Menard (2000), the following calculation is used:

$$\lambda_p = 1 - \frac{n - \sum_i f_{ii}}{n - n_{\text{mode}}} = \frac{\sum_i f_{ii} - n_{\text{mode}}}{n - n_{\text{mode}}},$$

where n_{mode} is the frequency of observed responses in the modal category of the outcome (maximum row marginal). For these data, $\lambda_p = .1739$ with the observed categories treated as the dependent variable. For the model constructed in the above example, predicting proficiency category membership (0, 1 versus 5) based on *gender* reduces the prediction error by 17.4%, once the marginal distribution of the DV is taken into account.

SAS produces several ordinal measures of association within the LOGISTIC procedure that can supplement the pseudo R^2's and the statistics for predictive efficiency determined from the classification table, such as Somers' D, a rank order correlation statistic (Cliff, 1996a; Liebetrau, 1983). Most of the rank order statistics are based on the notion of concordant versus discordant pairs. The term "pair" refers to pairing of each case (individual) with every other case in the data set (not including itself). For a sample of size n, there are $n(n-1)/2$ possible pairings of individuals. Of interest are pairs of individuals that do not have the same observed response; we ignore pairings for which both cases are 0 or both cases are 1 on the outcome of interest. If the two cases have dissimilar responses, the pair is called concordant when the predicted probability (of being classified as "success" based on the model) for the case with the observed value of 1 is higher than the case with the observed value of 0; otherwise, the pair is termed discordant. A pair (with dissimilar response) is tied if it cannot be classified as either concordant or discordant (this would happen if the predicted probabilities were very close; SAS categorizes predicted probabilities into interval lengths of .002 (SAS, 1997). The effect is to count the number of times the direction of prediction is accurate for each pair of individuals with different outcomes. Somers' D is probably the most widely used of the available rank order correlation statistics: Somers' D = (nc – nd)/t; where nc = number of concordant pairs, nd = number of discordant pairs, and t = number of pairs with different responses. Using SAS, Somers' D for this example is .189, which represents the strength of the correspondence between observed outcomes and predicted probabilities.[8]

Comparing Results Across Statistical Programs

To facilitate use and interpretation of logistic analysis across different statistical packages, as well as to lead into our discussion of the treatment of ordinal outcomes, the previous model was also fit using SAS PROC LOGISTIC (both descending and ascending approaches) and SPSS PLUM (for ordinal outcomes). A summary of results is shown in Table 3.2 (syntax for these models appears in Appendix A, sections A1–A4). All models used the logit link function.

TABLE 3.2

Comparison of Results for SPSS, SAS, and SPSS PLUM for a Dichotomous Outcome: Proficiency (0, 1 versus 5)[a] by Gender, $N = 702$

	SPSS Logistic and SAS (descending)	*SAS (ascending)*	*SPSS PLUM*
Probability estimated	$P(Y=1)$	$P(Y=0)$	$P(Y \le 0)$
Intercept	.430	–.430	.335
gender = 1 (male)	–.765**	.765**	0
gender = 0 (female)			.765**
Model fit			
–2LL (intercept only)	972.94	972.974	972.974[b]
–2LL (model)	947.825	947.825	947.825
χ^2_1 (p)	25.149 (< .0001)	25.149 (< .0001)	25.149 (< .0001)
Model predictions ($\hat{p}$)			
Male	.417	.583	.583
Female	.606	.394	.394

a. $Y = 0$ if response proficiency is 0 or 1; $Y = 1$ if response proficiency is 5.
b. Use "kernel" in the print command for SPSS PLUM to request the full value of the likelihoods.
$^{**}p < .01$.

Reviewing the results in the first column of Table 3.2, note that SPSS LOGISTIC REGRESSION and SAS PROC LOGISTIC (descending) are fitting the same model based around estimating $P(Y = 1)$, which is the probability that a child has a response in proficiency category 5. The probability predictions for these two identical models can be found by first calculating the logit for boys and girls using the estimates provided, exponentiating these logits to determine the odds for each group, and then transforming these odds back into probability for the response identified as "success" (p = [odds(success)/(1 + odds(success))].

The results for the second model, shown in Column 3 of Table 3.2, using SAS with the ascending option, simply model the probability that a child has a response in proficiency category 0 or 1, rather than the probability that a child has a response in category 5. Notice that the signs on the intercept and the effect for *gender* are reversed from those in Column 2, yet they are of the same magnitude. Also note that the sum of the probability estimates for boys in Columns 2 and 3 is equal to 1.0, and similarly for girls. SAS with the descending option (Column 2) models the complement of the event from the default approach (ascending, in Column 3). Thus, the probabilities derived from the ascending approach are the complementary probabilities to those found when using SAS with the descending option.

The model parameter estimates using SPSS PLUM look very different from those obtained using the earlier approaches, but in fact the probability estimates are identical to those in Column 3 (and therefore, by the rule of complements, can be used to find the probability estimates in Column 2). SPSS PLUM is a program specifically designed for analyzing ordinal response variables, and the resulting parameter estimates will not exactly correspond to those found under SPSS Logistic Regression. In particular, the probability being estimated in SPSS PLUM is the probability of a response being *at or below* a particular outcome value, that is, the lower category codes; in contrast, SPSS LOGISTIC models the probability of the category with the higher outcome value. Additionally, whereas SAS handles both dichotomous and ordinal responses through its LOGISTIC procedure, the SPSS PLUM procedure uses a slightly different formulation of the generalized linear model that looks like: $\ln(Y_j') = \theta_j - \beta_1 X_1$. In this expression, the subscript j refers to the response category, and X_1 refers to the single independent variable, *gender*. The estimate for the effect of *gender* is *subtracted* from the intercept. Another important distinction between PLUM results and those from logistic regression programs under SPSS or SAS is that PLUM internally sets up the coding for categorical predictors. In Column 4, the estimate provided for the *gender* effect corresponds to when *gender* = 0, that is, for females. The coding system used is clearly displayed on the printout (examples of PLUM and SAS printouts for ordinal models will be included in the next chapters). To find the estimated probability for a girl being (at most) in proficiency categories 0 or 1—that is, $P(Y \leq 0)$—which is equivalent in this case to $P(Y = 0)$ because there are no responses less than 0, we use the estimates to find the predicted logit for girls (.335 − .765 = −.43), exponentiate the result to find the odds for girls of being at (or below) $Y = 0$ (exp(−.43) = .65), and then solve for the estimated probability (.65/(1 + .65) = .394). The same process is used to find the estimated probability for boys of being at (or below) categories 0 or 1, or $P(Y = 0)$.

SPSS PLUM provides R^2_L, referred to as *McFadden's pseudo* R^2 (Long, 1997; Menard, 2000), in addition to R^2_{CS} and R^2_N. In order to obtain the necessary values for the −2LL deviance statistics, the "*kernel*" option must be specified in the SPSS PLUM "*/print*" statement, as shown in syntax A4 in Appendix A.

The previous discussion and simple comparison of how SAS and SPSS treat binary outcomes illustrate that although model parameter estimates may vary on the surface, the resulting predicted probabilities computed from the model estimates, as well as model fit statistics, are consistent across packages and approaches. These simple examples also illustrate that it is important for an analyst to be aware of the outcome being predicted as well as how categorical independent variables are incorporated into the models, once a statistical package is selected. Distinctions across approaches and packages become even more critical as the number of categories for an ordinal response variable increases beyond the binary case.

4. THE CUMULATIVE (PROPORTIONAL) ODDS MODEL FOR ORDINAL OUTCOMES

Overview of the Cumulative Odds Model

With only two categories for an outcome variable, logistic regression is used to model the likelihood of one of the outcomes, usually termed the "success," as a function of a set of independent variables. The estimated probabilities for the response of interest, *P*(success), as well as for its complement, 1 − *P*(success), can be determined using the prediction model for the logits, as shown in the example in Chapter 3. When the possible responses for an outcome variable consist of more than two categories and are ordinal in nature, the notion of "success" can be conceived of in many different ways. Regression models for ordinal response variables are designed for just this situation and are extensions of the logistic regression model for dichotomous data. The complexity in fitting ordinal regression models arises in part because there are so many different possibilities for how "success," and the consequent probability of "success," might be modeled.

For example, given a *K*-level ordinal response variable, such as proficiency in early literacy with $K = 6$ as in the ECLS-K study (Table 2.1), we could derive several different representations of "success" depending on how we view the data. In general, *K*-level ordinal data can be partitioned by $K - 1$ "success" cutpoints (Fox, 1997; McCullagh & Nelder, 1983). Success