Analysis of Ordinal
Categorical Data

Analysis of Ordinal Categorical Data

Second Edition

Alan Agresti

University of Florida
Gainesville, Florida

For general information on our other products and services or for technical support, please contact our
Customer Care Department within the United States at (800) 762-2974, outside the United States at
(317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print
may not be available in electronic formats. For more information about Wiley products, visit our web
site at www.wiley.com.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

In recent years methods for analyzing categorical data have matured considerably in their development. There has been a tremendous increase in the publication of research articles on this topic. Several books on categorical data analysis have introduced the methods to audiences of nonstatisticians as well as to statisticians, and the methods are now used frequently by researchers in areas as diverse as sociology, public health, and wildlife ecology. Yet some types of methods are still in the process of development, such as methods for clustered data, Bayesian methods, and methods for sparse data sets with large numbers of variables.

What distinguishes this book from others on categorical data analysis is its emphasis on methods for response variables having ordered categories, that is, *ordinal* variables. Specialized models and descriptive measures are discussed that use the information on ordering efficiently. These ordinal methods make possible simpler description of the data and permit more powerful inferences about population characteristics than do models for *nominal* variables that ignore the ordering information.

This is the second edition of a book published originally in 1984. At that time many statisticians were unfamiliar with the relatively new modeling methods for categorical data analysis, so the early chapters of the first edition introduced generalized linear modeling topics such as logistic regression and loglinear models. Since many books now provide this information, this second edition takes a different approach, assuming that the reader already has some familiarity with the basic methods of categorical data analysis. These methods include descriptive summaries using odds ratios, inferential methods including chi-squared tests of the hypotheses of independence and conditional independence, and logistic regression modeling, such as presented in Chapters 1 to 6 of my books *An Introduction to Categorical Data Analysis* (2nd ed., Wiley, 2007) and *Categorical Data Analysis* (2nd ed., Wiley, 2002).

On an ordinal scale, the technical level of this book is intended to fall between that of the two books just mentioned. I intend the book to be accessible to a broad audience, particularly professional statisticians and methodologists in areas such as public health, the pharmaceutical industry, the social and behavioral sciences, and business and government. Although there is some discussion of the underlying theory, the main emphasis is on presenting various ordinal methodologies. Thus, the book has more discussion of interpretation and application of the methods than

of the technical details. However, I also intend the book to be useful to specialists who may want to become aware of recent research advances, to supplement the background provided. For this purpose, the *Notes* section at the end of each chapter provides supplementary technical comments and embellishments, with emphasis on references to related research literature.

The text contains significant changes from and additions to the first edition, so it seemed as if I were writing a new book! As mentioned, the basic introductions to logistic regression and loglinear models have been removed. New material includes chapters on marginal models and random effects models for clustered data (Chapters 9 and 10) and Bayesian methods (Chapter 11), coverage of additional models such as the stereotype model, global odds ratio models, and generalizations of cumulative logit models, coverage of order-restricted inference, and more detail throughout on established methods.

Nearly all the methods presented can be implemented using standard statistical software packages such as R and S-Plus, SAS, SPSS, and Stata. The use of software for ordinal methods is discussed in the Appendix. The web site www.stat.ufl. edu/~aa/cda/software.html gives further details about software for applying methods of categorical data analysis. The web site www.stat.ufl.edu/~aa/ordinal/ord.html displays data sets not shown fully in the text (in the form of SAS programs), several examples of the use of a R function (mph.fit) that can conduct many of the nonstandard analyses in the text, and a list of known errata in the text.

The first edition was prepared mainly while I was visiting Imperial College, London, on sabbatical leave in 1981–1982. I would like to thank all who commented on the manuscript for that edition, especially Sir David Cox and Bent Jørgensen.

For this edition, special thanks to Maria Kateri and Joseph Lang for reading a complete draft and making helpful suggestions and critical comments. Maria Kateri also very generously provided bibliographic checking and pointed out many relevant articles that I did not know about. Thanks to Euijung Ryu for computing help with a few examples, for help with improving a graphic and with my LaTeX code, and for many helpful suggestions on the text and the Bibliography. Bhramar Mukherjee very helpfully discussed Bayesian methods for ordinal data and case–control methods and provided many suggestions about Chapter 11. Also, Ivy Liu and Bernhard Klingenberg made helpful suggestions based on an early draft, Arne Bathke suggested relevant research on rank-based methods, Edgar Brunner provided several helpful comments about rank-based methods and elegant ways of constructing statistics, and Carla Rampichini suggested relevant research on ordinal multilevel models. Thanks to Stu Lipsitz for data for Example 9.2.3 and to John Williamson and Kyungmann Kim for data for Example 9.1.3. Thanks to Beka Steorts for WinBUGS help, Cyrus Mehta for the use of StatXact, Jill Rietema for arranging for the use of SPSS, and Oliver Schabenberger for arranging for the use of SAS. I would like to thank co-authors of mine on various articles for permission to use materials from those articles. Finally, thanks as always to my wife, Jacki Levine, for her unwavering support during the writing of this book.

A truly wonderful reward of my career as a university professor has been the opportunity to work on research projects with Ph.D. students in statistics and with statisticians around the world. It is to them that I would like to dedicate this book.

ALAN AGRESTI

Gainesville, Florida and Brookline, Massachusetts
January 2010

CHAPTER 1

Introduction

1.1 ORDINAL CATEGORICAL SCALES

Until the early 1960s, statistical methods for the analysis of categorical data were at a relatively primitive stage of development. Since then, methods have been developed more fully, and the field of categorical data analysis is now quite mature. Since about 1980 there has been increasing emphasis on having data analyses distinguish between ordered and unordered scales for the categories. A variable with an ordered categorical scale is called *ordinal*. In this book we summarize the primary methods that can be used, and usually should be used, when response variables are ordinal.

Examples of ordinal variables and their ordered categorical scales (in parentheses) are opinion about government spending on the environment (too high, about right, too low), educational attainment (grammar school, high school, college, postgraduate), diagnostic rating based on a mammogram to detect breast cancer (definitely normal, probably normal, equivocal, probably abnormal, definitely abnormal), and quality of life in terms of the frequency of going out to have fun (never, rarely, occasionally, often). A variable with an unordered categorical scale is called *nominal*. Examples of nominal variables are religious affiliation (Protestant, Catholic, Jewish, Muslim, other), marital status (married, divorced, widowed, never married), favorite type of music (classical, folk, jazz, rock, other), and preferred place to shop (downtown, Internet, suburban mall). Distinct levels of such variables differ in quality, not in quantity. Therefore, the listing order of the categories of a nominal variable should not affect the statistical analysis.

Ordinal scales are pervasive in the social sciences for measuring attitudes and opinions. For example, each subject could be asked to respond to a statement such as "Same-sex marriage should be legal" using categories such as (strongly disagree, disagree, undecided, agree, strongly agree) or (oppose strongly, oppose

mildly, neutral, favor mildly, favor strongly). Such a scale with a neutral middle category is often called a *Likert scale*. Ordinal scales also occur commonly in medical and public health disciplines: for example, for variables describing pain (none, mild, discomforting, distressing, intense, excruciating), severity of an injury in an automobile crash (uninjured, mild injury, moderate injury, severe injury, death), illness after a period of treatment (much worse, a bit worse, the same, a bit better, much better), stages of a disease (I, II, III), and degree of exposure to a harmful substance, such as measuring cigarette smoking with the categories (nonsmoker, <1 pack a day, ≥1 pack a day) or measuring alcohol consumption of college students with the scale (abstainer, non-binge drinker, occasional binge drinker, frequent binge drinker). In all fields, ordinal scales result when inherently continuous variables are measured or summarized by researchers by collapsing the possible values into a set of categories. Examples are age measured in years (0–20, 21–40, 41–60, 61–80, above 80), body mass index (BMI) measured as (<18.5, 18.5–24.9, 25–29.9, ≥30) for (underweight, normal weight, overweight, obese), and systolic blood pressure measured as (<120, 120–139, 140–159, ≥160) for (normal, prehypertension, stage 1 hypertension, stage 2 hypertension).

Often, for each observation the choice of a category is subjective, such as in a subject's report of pain or in a physician's evaluation regarding a patient's stage of a disease. (An early example of such subjectivity was U.S. President Thomas Jefferson's suggestion during his second term that newspaper articles could be classified as truths, probabilities, possibilities, or lies.) To lessen the subjectivity, it is helpful to provide guidance about what the categories represent. For example, the College Alcohol Study conducted at the Harvard School of Public Health defines "binge drinking" to mean at least five drinks for a man or four drinks for a woman within a two-hour period (corresponding to a blood alcohol concentration of about 0.08%); "occasional binge drinking" is defined as binge drinking once or twice in the past two weeks; and "frequent binge drinking" is binge drinking at least three times in the past two weeks.

For ordinal scales, unlike *interval* scales, there is a clear ordering of the levels, but the absolute distances among them are unknown. Pain measured with categories (none, mild, discomforting, distressing, intense, excruciating) is ordinal, because a person who chooses "mild" feels more pain than if he or she chose "none," but no numerical measure is given of the difference between those levels. An ordinal variable is *quantitative*, however, in the sense that each level on its scale refers to a greater or smaller magnitude of a certain characteristic than another level. Such variables are of quite a different nature than qualitative variables, which are measured on a nominal scale and have categories that do not relate to different magnitudes of a characteristic.

1.2   ADVANTAGES OF USING ORDINAL METHODS

Many well-known statistical methods for categorical data treat all response variables as nominal. That is, the results are invariant to permutations of the categories

of those variables, so they do not utilize the ordering if there is one. Examples are the Pearson chi-squared test of independence and multinomial response modeling using baseline-category logits. Test statistics and P-values take the same values regardless of the order in which categories are listed. Some researchers routinely apply such methods to nominal and ordinal variables alike because they are both categorical.

Recognizing the discrete nature of categorical data is useful for formulating sampling models, such as in assuming that the response variable has a multinomial distribution rather than a normal distribution. However, the distinction regarding whether data are continuous or discrete is often less crucial to substantive conclusions than whether the data are qualitative (nominal) or quantitative (ordinal or interval). Since ordinal variables are inherently quantitative, many of their descriptive measures are more like those for interval variables than those for nominal variables. The models and measures of association for ordinal data presented in this book bear many resemblances to those for continuous variables.

A major theme of this book is how to analyze ordinal data by utilizing their quantitative nature. Several examples show that the type of ordinal method used is not that crucial, in the sense that we obtain similar substantive results with ordinal logistic regression models, loglinear models, models with other types of response functions, or measures of association and nonparametric procedures. These results may be quite different, however, from those obtained using methods that treat all the variables as nominal.

Many advantages can be gained from treating an ordered categorical variable as ordinal rather than nominal. They include:

- Ordinal data description can use measures that are similar to those used in ordinary regression and analysis of variance for quantitative variables, such as correlations, slopes, and means.
- Ordinal analyses can use a greater variety of models, and those models are more parsimonious and have simpler interpretations than the standard models for nominal variables, such as baseline-category logit models.
- Ordinal methods have greater power for detecting relevant trend or location alternatives to the null hypothesis of "no effect" of an explanatory variable on the response variable.
- Interesting ordinal models apply in settings for which standard nominal models are trivial or else have too many parameters to be tested for goodness of fit.

An ordinal analysis can give quite different and much more powerful results than an analysis that ignores the ordinality. For a preview of this, consider Table 1.1, with artificial counts in a contingency table designed to show somewhat of a trend from the top left corner to the bottom right corner. For two-way contingency tables, the first analysis many methodologists apply is the chi-squared test of independence. The Pearson statistic equals 10.6 with df = 9, yielding an unimpressive P-value of 0.30. By contrast, various possible ordinal analyses for testing this hypothesis have

TABLE 1.1. Data Set for Which Ordinal Analyses Give Very Different Results from Unordered Categorical Analyses

	Column 1	Column 2	Column 3	Column 4
Row 1	8	6	4	2
Row 2	6	8	6	4
Row 3	4	6	8	6
Row 4	2	4	6	8

chi-squared statistics on the order of 9 or 10, but with df = 1, and have P-values on the order of 0.002 and 0.001.

1.3 ORDINAL MODELING VERSUS ORDINARY REGESSION ANALYSIS

There are two relatively extreme ways to analyze ordered categorical response variables. One way, still common in practice, ignores the categorical nature of the response variable and uses standard parametric methods for continuous response variables. This approach assigns numerical scores to the ordered categories and then uses ordinary least squares (OLS) methods such as linear regression and analysis of variance (ANOVA). The second way restricts analyses solely to methods that use only the ordering information about the categories. Examples of this approach are nonparametric methods based on ranks and models for cumulative response probabilities.

1.3.1 Latent Variable Models for Ordinal Data

Many other methods fall between the two extremes described above, using ordinal information but having some parametric structure as well. For example, often it is natural to assume that an unobserved continuous variable underlies the ordinal response variable. Such a variable is called a *latent variable*.

In a study of political ideology, for example, one survey might use the categories liberal, moderate, and conservative, whereas another might use very liberal, slightly liberal, moderate, slightly conservative, and very conservative or an even finer categorization. We could regard such scales as categorizations of an inherently continuous scale that we are unable to observe. Then, rather than assigning scores to the categories and using ordinary regression, it is often more sensible to base description and inference on parametric models for the latent variable. In fact, we present connections between this approach and a popular modeling approach that has strict ordinal treatment of the response variable: In Chapters 3 and 5 we show that a logistic model and a probit model for cumulative probabilities of an ordinal response variable can be motivated by a latent variable model for an underlying quantitative response variable that has a parametric distribution such as the normal.

1.3.2 Using OLS Regression with an Ordinal Response Variable

In this book we do present methods that use only the ordering information. It is often attractive to begin a statistical analysis by making as few assumptions as possible, and a strictly ordinal approach does this. However, in this book we also present methods that have some parametric structure or that require assigning scores to categories. We believe that strict adherence to operations that utilize only the ordering in ordinal scales limits the scope of useful methodology too severely. For example, to utilize the ordering of categories of an ordinal explanatory variable, nearly all models assign scores to the categories and regard the variable as quantitative—the alternative being to ignore the ordering and treat the variable as nominal, with indicator variables. Therefore, we do not take a rigid view about permissible methodology for ordinal variables.

That being said, we recommend against the simplistic approach of posing linear regression models for ordinal response scores and fitting them using OLS methods. Although that approach can be useful for identifying variables that clearly affect a response variable, and for simple descriptions, limitations occur. First, there is usually not a clear-cut choice for the scores. Second, a particular response outcome is likely to be consistent with a range of values for some underlying latent variable, and an ordinary regression analysis does not allow for the measurement error that results from replacing such a range by a single numerical value. Third, unlike the methods presented in this book, that approach does not yield estimated probabilities for the response categories at fixed settings of the explanatory variables. Fourth, that approach can yield predicted values above the highest category score or below the lowest. Fifth, that approach ignores the fact that the variability of the responses is naturally nonconstant for categorical data: For an ordinal response variable, there is little variability at predictor values for which observations fall mainly in the highest category (or mainly in the lowest category), but there is considerable variability at predictor values for which observations tend to be spread among the categories.

Related to the second, fourth, and fifth limitations, the ordinary regression approach does not account for "ceiling effects" and "floor effects," which occur because of the upper and lower limits for the ordinal response variable. Such effects can cause ordinary regression modeling to give misleading results. These effects also result in substantial correlation between values of residuals and values of quantitative explanatory variables.

1.3.3 Example: Floor Effect Causes Misleading OLS Regression

How can ordinary regression give misleading results when used with ordered categorical response variables? To illustrate, we apply the standard linear regression model to simulated data with an ordered categorical response variable y based on an underlying continuous latent variable y^*. The explanatory variables are a continuous variable x and a binary variable z. The data set of 100 observations was generated as follows: The x values were independently uniformly generated between 0 and 100, and the z values were independently generated with $P(z = 0) = P(z = 1) = 0.50$. At a given x, the latent response outcome y^* was generated according to a normal

distribution with mean

$$E(y^*) = 20.0 + 0.6x - 40.0z$$

and standard deviation 10. The first scatterplot in Figure 1.1 shows the 100 observations on y^* and x, each data point labeled by the category for z. The plot also shows the OLS fit that estimates this model.

We then categorized the 100 generated values on y^* into five categories to create observations for an ordinal variable y, as follows:

$$y = 1 \text{ if } y^* \le 20, \quad y = 2 \text{ if } 20 < y^* \le 40, \quad y = 3 \text{ if } 40 < y^* \le 60,$$

$$y = 4 \text{ if } 60 < y^* \le 80, \quad y = 5 \text{ if } y^* > 80.$$

The second scatterplot in Figure 1.1 shows 100 observations on y and x. At low x levels, there is a floor effect for the observations with $z = 1$. When $x < 50$ with $z = 1$, there is a very high probability that observations fall in the lowest category of y.

Using OLS with scores 1, 2, 3, 4, and 5 for the categories of y suggests either (a) a model with an interaction term, allowing different slopes relating $E(y)$ to x when
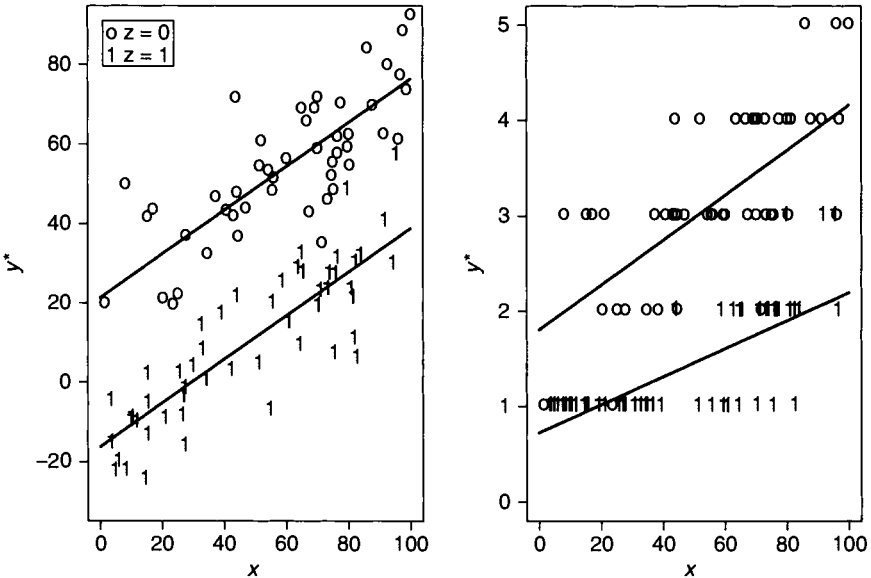
Figure 1.1. Ordered categorical data (in second panel) for which ordinary regression suggests interaction, because of a floor effect, but ordinal modeling does not. The data were generated (in first panel) from a normal main-effects regression model with continuous (x) and binary (z) explanatory variables. When the continuous response y^* is categorized and y is measured as (1, 2, 3, 4, 5), the observations labeled "1" for the category of z have a linear x effect with only half the slope of the observations labeled "0" for the category of z.

$z = 0$ and when $z = 1$, or (b) a model with a quadratic effect of x on $E(y)$ when $z = 1$. The second scatterplot in Figure 1.1 shows the fit of the linear interaction model, that is, using OLS to fit the model $E(y) = \alpha + \beta_1 x + \beta_2 z + \beta_3 (x \times z)$ to the ordered categorical response. The slope of the line is about twice as high when $z = 0$ as when $z = 1$. This interaction effect is caused by the observations when $z = 1$ tending to fall in category $y = 1$ whenever x takes a relatively low value. As x gets lower, the underlying value y^* can continue to tend to get lower, but the observed ordinal response cannot fall below 1.

Standard ordinal models such as those introduced in Chapters 3 to 5 fit the data well without the need for an interaction term. Such models can be motivated by a latent variable model. They allow for underlying values of y^* when $z = 1$ to be below those when $z = 0$, even if x is so low that y is very likely to be in the first category at both levels of z. (The data in Figure 1.1 are revisited with such a model in Exercise 5.2.)

Hastie et al. (1989) showed a real-data example of the type we presented here with simulated data. They described a study of women in South Africa that modeled an ordinal measurement y of osteoporosis in terms of x = age and an indicator variable z for whether the woman had osteoarthritis. At low age levels, a high proportion of women clustered in the lowest category of osteoporosis, regardless of osteoarthritis status. Using OLS, for each osteoarthritis group the line relating age to the predicted osteoporosis score took value at the lowest ordinal level near a relatively low age level, but the line for the group positive for osteoarthritis had a significantly greater slope as age increased. In fact, there was also a significant quadratic effect for that group. When the authors used an ordinal model instead, they found no evidence of interaction. For other such examples, see McKelvey and Zavoina (1975, Sec. 4) and Winship and Mare (1984).

1.3.4 Ordinal Methods with Truly Quantitative Data

Even when the response variable is interval scale rather than ordered categorical, ordinal models can still be useful. One such case occurs when the response outcome is a count but when standard sampling models for counts, such as the Poisson, do not apply. For example, each year the British Social Attitudes Survey asks a sample of people their opinions on a wide range of issues. In several years the survey asked whether abortion should be legal in each of seven situations, such as when a woman is pregnant as a result of rape. The number of cases to which a person responds "yes" is a summary measure of support for legalized abortion. This response variable takes values between 0 and 7. It is inappropriate to treat it as a binomial variate because the separate situations would not have the same probability of a "yes" response or have independent responses. It is inappropriate to treat it as a Poisson or negative binomial variate, because there is an upper bound for the possible outcome, and at some settings of explanatory variables most observations could cluster at the upper limit of 7. Methods for ordinal data *are* valid, treating each observation as a single multinomial trial with eight ordered categories.

For historical purposes it is interesting to read the extensive literature of about 40 years ago, much of it in the social sciences, regarding whether it is permissible to assign scores to ordered categories and use ordinary regression methods. See, for example, Borgatta (1968), Labovitz (1970), and Kim (1975) for arguments in favor and Hawkes (1971), Mayer (1971), and Mayer and Robinson (1978) for arguments against.

1.4 ORGANIZATION OF THIS BOOK

The primary methodological emphasis in this book is on models that describe associations and interactions and provide a framework for making inferences. In Chapter 2 we introduce ordinal odds ratios that are natural parameters for describing most of these models. In Chapter 3 we introduce the book's main focus, presenting logistic regression models for the cumulative probabilities of an ordinal response. In Chapter 4 we summarize other types of models that apply a logit link function to ordinal response variables, and in Chapter 5 we present other types of link functions for such models.

The remainder of the book deals with multivariate ordinal responses. In Chapter 6 we present loglinear and other models for describing association and interaction structure among a set of ordinal response variables, and in Chapter 7 present bivariate ordinal measures of association that summarize the entire structure by a single number. The following three chapters deal with multivariate ordinal responses in which each response has the same categories, such as happens in longitudinal studies and other studies with repeated measurement. This topic begins in Chapter 8 with methods for square contingency tables having ordered rows and the same ordered columns and considers applications in which such tables arise. Chapters 9 and 10 extend this to an analysis of more general forms of correlated, clustered ordinal responses. Primary attention focuses on models for the marginal components of a multivariate response and on models with random effects for the clusters.

In Chapters 2 to 10 we take a frequentist approach to statistical inference, focusing on methods that use only the likelihood function. In the final chapter we show ways of implementing Bayesian methods with ordinal response variables, combining prior information about the parameters with the likelihood function to obtain a posterior distribution of the parameters for inference. The book concludes with an overview of software for the analysis of ordered categorical data, emphasizing R and SAS.

For other surveys of methods for ordinal data, see Hildebrand et al. (1977), Agresti (1983a, 1999), Winship and Mare (1984), Armstrong and Sloan (1989), Barnhart and Sampson (1994), Clogg and Shihadeh (1994), Ishii-Kuntz (1994), Ananth and Kleinbaum (1997), Scott et al. (1997), Johnson and Albert (1999), Bender and Benner (2000), Guisan and Harrell (2000), Agresti and Natarajan (2001), Borooah (2002), Cliff and Keats (2002), Lall et al. (2002), Liu and Agresti (2005), and O'Connell (2006).

CHAPTER 2

Ordinal Probabilities, Scores, and Odds Ratios

In this chapter we introduce ways of using odds ratios and other summary measures to describe the association between two ordinal categorical variables. The measures apply to sample data or to a population. We also present confidence intervals for these measures. First, though, we introduce some probabilities and scores that are a basis of ways of describing marginal and conditional distributions of ordinal response variables.

2.1 PROBABILITIES AND SCORES FOR AN ORDERED CATEGORICAL SCALE

For an ordinal response variable Y, let c denote the number of categories. For n observations in a sample, $n_1, n_2, \ldots, n_c$ denote the frequencies in the categories, with $n = \sum_j n_j$, and $\{p_j = n_j/n\}$ denote the sample proportions.

For an observation randomly selected from the corresponding population, let π_j denote the probability of response in category j. Some measures and some models utilize the *cumulative probabilities*

$$F_j = P(Y \leq j) = \pi_1 + \cdots + \pi_j, \qquad j = 1, 2, \ldots, c.$$

These reflect the ordering of the categories, with

$$0 < F_1 < F_2 < \cdots < F_c = 1.$$

2.1.1 Types of Scores for Ordered Categories

How can summary measures utilize the ordinal nature of the categorical scale? One simple way uses the cumulative probabilities to identify the *median* response:

namely, the minimum j such that $F_j \geq 0.50$. With a categorical response, an unappealing aspect of this measure for making comparisons of groups is its discontinuous nature: Changing a tiny bit of probability can have the effect of moving the median from one category to the next. Also, two groups can have the same median even when an underlying latent variable has distribution shifted upward for one group relative to the other.

Alternatively, we could assign ordered scores

$$v_1 < v_2 < \cdots < v_c$$

to the categories and summarize the observations with ordinary measures for quantitative data such as the *mean*. Doing this treats the ordinal scale as an interval scale. There is no unique way to select scores, and the key aspect is the choice for the relative distances between pairs of adjacent categories. For example, with $c = 3$, comparisons of means for two groups using the scores (1, 2, 3) yields the same substantive conclusions as using the scores (0, 5, 10) or any set of linearly transformed scores but possibly different conclusions from using scores such as (1, 2, 5) or (0, 3, 10). Often, an appropriate choice of scores is unclear. In that case it is advisable to perform a sensitivity analysis: Choose scores in a few sensible ways that are not linear translations, and check whether conclusions for the method that uses those scores depend on the choice.

An alternative approach to selecting scores uses the data themselves to determine the scores. One such set uses the average cumulative proportions for the ordinal response variable. For sample proportions $\{p_j\}$, the average cumulative proportion in category j is

$$a_j = \sum_{k=1}^{j-1} p_k + \frac{1}{2}p_j, \qquad j = 1, 2, \ldots, c,$$

that is, the proportion of subjects below category j plus half the proportion in category j. In terms of the sample cumulative proportions $\hat{F}_j = p_1 + \cdots + p_j$,

$$a_j = \frac{\hat{F}_{j-1} + \hat{F}_j}{2},$$

with $\hat{F}_0 = 0$. Bross (1958) introduced the term *ridits* for the average cumulative proportion scores.

The ridits have the same ordering as the categories, $a_1 \leq a_2 \leq \cdots \leq a_c$. Their weighted average with respect to the sample distribution satisfies

$$\sum_{j=1}^{c} p_j a_j = \sum_{j=1}^{c} p_j \left(\sum_{k=1}^{j-1} p_k + \frac{1}{2}p_j \right)$$

$$= \frac{2 \sum \sum_{k<j} p_j p_k + \sum_j p_j^2}{2} = \frac{\left(\sum_j p_j \right)^2}{2} = 0.50.$$

The ridits are linearly related to the *midranks*, which are the averages of the ranks that would be assigned if the observations in a category could be ranked without ties. The midrank r_1 for category 1 is the average of the ranks $1, \ldots, n_1$ that pertain to the n_1 observations in category 1, so $r_1 = (1 + n_1)/2$. The midrank for category 2 is $r_2 = [(n_1 + 1) + (n_1 + n_2)]/2$. Generally, the midrank for category j is

$$r_j = \frac{[(\sum_{i=1}^{j-1} n_i) + 1] + \sum_{i=1}^{j} n_i}{2}.$$

Whereas midrank scores fall between 1 and n, ridit scores fall between 0 and 1. The linear relationship between them is

$$r_j = na_j + 0.5, \qquad a_j = \frac{r_j - 0.5}{n}.$$

Ridit and midrank scores take directly into account the way the response is categorized. For example, if two adjacent categories are combined, the ridit (or midrank) score for the new category falls between the original two scores, with the other scores being unaffected. If the category ordering is reversed, the ridit score for category j transforms from a_j to $1 - a_j$.

Another way to form data-dependent scores assumes a particular distribution for an unobserved continuous *latent variable* assumed to underlie Y. This approach regards the ordinal scale as representing a partition of intervals of values of the latent variable. For example, suppose that we assume an underlying standard normal distribution, with cumulative distribution function Φ. Then we could use some variation of *normal scores* as applied in some nonparametric statistical methods. For example, we could let v_1 be the mean of the truncated normal distribution falling between $-\infty$ and $\Phi^{-1}(p_1)$ [where $\Phi^{-1}(p_1)$ denotes the standard-normal score for which the cumulative probability below it equals p_1], let v_2 be the mean of the truncated normal distribution falling between $\Phi^{-1}(p_1)$ and $\Phi^{-1}(p_1 + p_2)$, and so on, up to v_c, which is the mean of the truncated normal distribution falling between $\Phi^{-1}(p_1 + \cdots + p_{c-1})$ and ∞. More simply, we could let $v_j = \Phi^{-1}(a_j)$ where a_j is the ridit score in category j. A very similar score based on the midranks $\{r_j\}$ is $v_j = \Phi^{-1}[r_j/(n + 1)]$.

We used scores in this section to summarize ordinal data, but it is not necessary to do so. In this chapter we learn about other methods that do not require assigning scores, and this is also true of most models for ordinal response variables presented in later chapters.

2.1.2 Example: Belief in Heaven

Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of adult Americans provides data about the opinions and behaviors of the American public. It is simple to download results from the surveys.[1] In this book we use several data sets from the GSS to illustrate methods.

[1] This can currently be done at sda.berkeley.edu/GSS.

TABLE 2.1. Responses About Belief in Heaven

	Does Heaven Exist?				
	Definitely	Probably	Probably Not	Definitely Not	Total
Count	1546	498	205	138	2387
Proportion	0.648	0.209	0.086	0.058	1.0
Ridit score	0.324	0.752	0.899	0.971	

Source: General Social Survey.

Table 2.1 shows results of 2387 responses from the GSS to a question about whether heaven exists. The ridit scores for the counts in this ordinal categorical scale are

$$a_1 = \left(\frac{1}{2}\right)\frac{1546}{2387} = 0.32, \qquad a_2 = \frac{1546}{2387} + \left(\frac{1}{2}\right)\frac{498}{2387} = 0.75,$$

$$a_3 = \frac{1546 + 498}{2387} + \left(\frac{1}{2}\right)\frac{205}{2387} = 0.90,$$

$$a_4 = \frac{1546 + 498 + 205}{2387} + \left(\frac{1}{2}\right)\frac{138}{2387} = 0.97.$$

The ridit scores of 0.90 for "probably not" and 0.97 for "definitely not" are relatively close. Whenever two adjacent categories both have relatively small proportions, this necessarily happens.

The normal scores based on ridits, $v_j = \Phi^{-1}(a_j)$, are $(-0.457, 0.681, 1.277, 1.897)$, where, for example, $\Phi(-0.457) = a_1 = 0.32$ is the probability that a standard normal variable falls below -0.457. The very similar normal scores based on midranks, $v_j = \Phi^{-1}[r_j/(n+1)]$, are $(-0.457, 0.680, 1.276, 1.894)$, where, for example, $\Phi(-0.457) = [(1 + 1546)/2]/2388 = 0.324$.

This example illustrates that ridit scores or scores based on them, such as normal scores, need not represent an underlying scale realistically. For the ridit scores (0.32, 0.75, 0.90, 0.97) for (definitely, probably, probably not, definitely not), the score of 0.75 for "probably" is closer to the score of 0.97 for "definitely not" than it is to the score of 0.32 for "definitely." Yet we would not be likely to regard "probably" and "definitely not" as closer together than "probably" and "definitely." Similarly, note that the normal scores treat "definitely" and "probably" as being nearly twice as far apart as "probably" and "probably not" or "probably not" and "definitely not."

For descriptive summaries of this ordinal scale, such as comparing mean responses for different groups, it is often more sensible to use fixed scores instead of ridit scores or normal scores. The scores (1, 2, 3, 4) would treat (definitely, probably, probably not, definitely not) as equidistant for pairs of adjacent categories. Scores such as (0, 1, 4, 5) would treat the distance between "probably" and "probably not" as greater than the distance between "definitely" and "probably" and the distance between "probably not" and "definitely not."

2.1.3 Two-Way Contingency Tables with an Ordinal Response

In practice, observations on ordinal response variables are usually accompanied by observations on explanatory variables and are sometimes accompanied by observations on other response variables. When the other variables are categorical, a contingency table can display the frequencies of observations for the various combinations of levels of the variables. Each cell in the contingency table shows the number of observations that have that combination. In this chapter we consider primarily the case of two categorical variables. We denote the second variable by X if it is another response variable and by x if it is an explanatory variable. We let r denote the number of rows and let c denote the number of columns in the contingency table. Let n_{ij} denote the number of observations in the cell of the table in row i and column j.

For a two-way cross-classification of an ordinal response variable Y with another categorical response variable X, let $\{p_{ij}\}$ denote the cell proportions for the possible values of (X, Y). That is, $p_{ij} = n_{ij}/n$, where n is the total sample size. Then $\sum_i \sum_j p_{ij} = 1$, and $\{p_{ij}\}$ is the sample *joint distribution*. The sample marginal distributions are the row totals and column totals obtained by summing the joint proportions. We denote marginal proportions by p_{i+} for row i and p_{+j} for column j. Note that $p_{+j} = \sum_i p_{ij} = \sum_i n_{ij}/n$ and $\sum_j p_{+j} = 1$.

Although the second variable could also be a response variable, more commonly it is an explanatory variable. Then *conditional distributions* for the response variable are usually more relevant than joint distributions. We let the columns refer to the ordinal response variable Y and the rows refer to the explanatory variable x. For the observations in row i, we denote the proportion in category j of Y by $p_{j|i}$. Hence, $p_{j|i} = n_{ij}/n_{i+}$, where n_{i+} is the total count in row i and $\sum_j p_{j|i} = 1$ for each i. The values $(p_{1|i}, p_{2|i}, \ldots, p_{c|i})$ form a sample conditional distribution. Different levels of x can be compared with respect to the proportions of observations in the various categories of Y. The sample conditional cumulative proportions,

$$\hat{F}_{j|i} = p_{1|i} + \cdots + p_{j|i}, \qquad j = 1, 2, \ldots, c,$$

specify the proportion of observations classified in one of the first j columns, given classification in row i.

2.1.4 Probabilistic Comparisons of Two Ordinal Distributions

Now consider the special case of a $2 \times c$ table, for comparing two groups on an ordinal response variable Y. Let Y_1 and Y_2 denote the column numbers of the response variable for subjects selected at random from rows 1 and 2, independent of each other. A measure that summarizes their relative size is

$$\alpha = P(Y_1 > Y_2) + \tfrac{1}{2} P(Y_1 = Y_2) \tag{2.1}$$

(Kruskal 1957; Klotz 1966). If Y_1 and Y_2 are identically distributed or if they have symmetric distributions over all c categories, then $\alpha = 0.50$. When $\alpha > 0.50$ (< 0.50), outcomes of Y_1 tend to be larger (smaller) than outcomes of Y_2.

A related measure that has null value equal to 0 rather than 0.50 is

$$\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1). \qquad (2.2)$$

The measures α and Δ are functionally related,

$$\alpha = \frac{\Delta + 1}{2}, \qquad \Delta = 2\alpha - 1,$$

with α having range $[0, 1]$ and Δ having range $[-1, 1]$. We refer to them as measures of *stochastic superiority*, a term introduced by Vargha and Delaney (1998). In Chapter 7 we present related measures for $r \times c$ tables.

With sample data we can estimate α from the conditional distributions by

$$\hat{\alpha} = \sum_{j > k} \sum p_{j|1} p_{k|2} + \tfrac{1}{2} \sum_j p_{j|1} p_{j|2}.$$

The sample version of Δ is

$$\hat{\Delta} = \sum_{j > k} \sum p_{j|1} p_{k|2} - \sum_{j < k} \sum p_{j|1} p_{k|2}.$$

Another useful comparison of $P(Y_1 > Y_2)$ and $P(Y_2 > Y_1)$ is

$$\theta = \frac{P(Y_1 > Y_2)}{P(Y_2 > Y_1)}.$$

Its sample value is

$$\hat{\theta} = \frac{\sum \sum_{j > k} p_{j|1} p_{k|2}}{\sum \sum_{j < k} p_{j|1} p_{k|2}} = \frac{\sum \sum_{j > k} n_{1j} n_{2k}}{\sum \sum_{j < k} n_{1j} n_{2k}}.$$

When $c = 2$, $\hat{\theta}$ is an odds ratio. For $c > 2$, $\hat{\theta}$ is a generalized odds ratio for ordinal responses (Agresti 1980), which we refer to as an *ordinal odds ratio* for comparing two groups. In Section 2.2 we introduce other ways of forming odds ratios for ordinal responses.

The ordinal odds ratio θ differs slightly from

$$\frac{\alpha}{1 - \alpha} = \frac{P(Y_1 > Y_2) + \tfrac{1}{2} P(Y_1 = Y_2)}{P(Y_2 > Y_1) + \tfrac{1}{2} P(Y_1 = Y_2)},$$

which approximates $P(Y_1 > Y_2)/P(Y_2 > Y_1)$ for an underlying continuous scale. The measure $\alpha/(1 - \alpha)$ is closer to 1.0 than is θ. Similarly, usually $P(Y_1 > Y_2)/P(Y_2 > Y_1)$ for the underlying continuous scale is closer to 1.0 than θ is for the observed ordinal scale. This is because observations that are tied on the observed ordinal scale usually have similar relative frequencies of the two orders for the

underlying scale. By contrast, we can interpret α or Δ either for the observed scale or an underlying continuous scale. For example, suppose that $\Delta = 0.40$. Then, in comparisons of the groups with independent observations for the underlying continuum, we expect a higher response for group 1 about 70% of the time and a higher response for group 2 about 30% of the time, since $0.70 - 0.30 = 0.40$ and $0.70 + 0.30 = 1.0$.

2.1.5   Means of Conditional Distributions in Two-Way Tables

Next we consider $r \times c$ tables. With ordered scores $\{v_j\}$ for the categories of Y, in each row we can use the sample conditional distribution to find a sample mean response. In row i this is

$$\bar{y}_i = \sum_{j=1}^{c} v_j p_{j|i}.$$

When x is ordinal, we often expect a trend (upward or downward) in $\{\bar{y}_i\}$ across the rows.

Alternatively, we could find the means using data-generated scores. For example, we could use ridit scores for Y calculated from the proportions in its marginal distribution. For outcome category j,

$$a_j = \sum_{k=1}^{j-1} p_{+k} + \tfrac{1}{2} p_{+j}, \qquad j = 1, 2, \ldots, c.$$

The *mean ridit* for the sample conditional distribution in row i is

$$\bar{A}_i = \sum_{j=1}^{c} a_j p_{j|i}.$$

The weighted average of the mean ridits satisfies

$$\sum_{i=1}^{r} p_{i+} \bar{A}_i = 0.50.$$

When the data in the full sample are ranked, using midranks $\{r_j\}$, the *mean rank* for the sample conditional distribution in row i is

$$\bar{R}_i = \sum_{j=1}^{c} r_j p_{j|i}.$$

Their weighted average over the r rows is $(n + 1)/2$. The mean ridits and mean ranks are related by

$$\bar{A}_i = \frac{\bar{R}_i - 0.50}{n}.$$

Bross (1958) argued that an advantage of ridit scoring is their lack of sensitivity to the way the ordinal response variable is categorized (e.g., with different numbers of categories). Two researchers who categorize an ordinal response in different ways for a particular sample would, nevertheless, obtain similar mean ridits for the rows.

2.1.6 Mean Ridits and Mean Ranks Relate to Stochastic Superiority Measures

For $2 \times c$ tables, the sample values of the stochastic superiority measures $\alpha = P(Y_1 > Y_2) + \frac{1}{2} P(Y_1 = Y_2)$ and $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$ relate to the mean ridit scores in the two rows by[2]

$$\hat{\alpha} = (\bar{A}_1 - \bar{A}_2 + 0.50) \quad \text{and} \quad \hat{\Delta} = 2(\bar{A}_1 - \bar{A}_2).$$

Vigderhous (1979) presented other connections between mean ridit measures and ordinal measures of association. In terms of the mean ranks $\bar{R}_1$ and $\bar{R}_2$ in the two rows,

$$\hat{\alpha} = \frac{\bar{R}_1 - \bar{R}_2}{n} + 0.50 \quad \text{and} \quad \hat{\Delta} = \frac{2(\bar{R}_1 - \bar{R}_2)}{n}.$$

For $r \times c$ tables, let Y_i denote the response outcome for a randomly selected subject at level i of x, and let Y^* denote the response outcome for a randomly selected subject from the marginal distribution of Y. The sample mean ridit $\bar{A}_i$ using the marginal ridit scores estimates

$$P(Y_i > Y^*) + \frac{1}{2} P(Y_i = Y^*).$$

In analogy with the terms *logit* and *probit*, Bross (1958) chose the term *ridit* because $\bar{A}_i$ describes how the distribution of Y in row i compares *r*elative to an *i*dentified *d*istribution (in this case, the marginal distribution of Y). The $\{\bar{A}_i\}$ or the corresponding population values can be used to compare each row to an overall marginal distribution of the response (Kruskal 1952). In some of the literature on nonparametric statistical methods they are referred to as *relative effects*.

For underlying continuous distributions, $\bar{A}_i$ estimates the probability that an observation from row i ranks higher on the ordinal response variable than does an observation from the marginal distribution of Y. Such a probability inference is approximate, since besides sampling error, it is unknown how tied observations for the observed discrete scale would be ordered for an underlying continuum. Also, the sample marginal distribution of Y, which determines the ridit scores, reflects the study design. For some sampling schemes, this need not be close to the population marginal distribution.

---

[2]For fully ranked data, analogous connections exist between Wilcoxon statistics using mean ranks and Mann–Whitney statistics using pairwise orderings.

To compare rows i and k in an $r \times c$ table, it is tempting to regard $(\bar{A}_i - \bar{A}_k + 0.50)$ as an estimate of $[P(Y_i > Y_k) + \frac{1}{2} P(Y_i = Y_k)]$, as suggested by Bross (1958). However, this may be a highly biased estimate, mainly because $P(Y_i > Y_k)$ is not determined by $P(Y_i > Y^*)$ and $P(Y_k > Y^*)$. To estimate $[P(Y_i > Y_k) + \frac{1}{2} P(Y_i = Y_k)]$, it is more appropriate to find $(\bar{A}_i - \bar{A}_k + 0.50)$ by computing the ridit scores using rows i and k alone (Beder and Heim 1990). This is equivalent to estimating α using data in that pair of rows alone (see also Note 2.2).

2.1.7 Example: Comparing Treatments for Gastric Ulcer Crater

We illustrate these methods for comparing two ordinal categorical distributions using Table 2.2 from a randomized study to compare two treatments for a gastric ulcer crater. The response, following three months of treatment, was the change in the size of the ulcer crater. This was measured with the ordinal scale (larger, less than $\frac{2}{3}$ healed, $\frac{2}{3}$ or more healed, healed).

The sample conditional distributions on the ordinal response are:

$$\text{Treatment A}: \quad (0.19, 0.12, 0.31, 0.38).$$

$$\text{Treatment B}: \quad (0.34, 0.25, 0.25, 0.16).$$

For the category scores $(1, 2, 3, 4)$, the sample mean responses are $\bar{y}_1 = 2.875$ and $\bar{y}_2 = 2.219$, indicating a tendency for a better response with treatment A. The sample ridit scores using the response marginal distribution are

$$a_1 = 0.133, \quad a_2 = 0.359, \quad a_3 = 0.594, \quad a_4 = 0.867.$$

The sample mean ridits are $\bar{A}_1 = 0.581$ and $\bar{A}_2 = 0.419$. Thus, $\hat{\alpha} = (\bar{A}_1 - \bar{A}_2 + 0.50) = 0.661$ estimates the probability of a better response with treatment A than treatment B, for underlying continuous responses. Similarly, $\hat{\Delta} = 2(\bar{A}_1 - \bar{A}_2) = 0.322$ estimates the difference between the probability that the response is better with A than B and the probability that the response is better with B than A for the observed scale or for an underlying continuous scale. The ordinal odds ratio for the observed scale is

$$\hat{\theta} = \frac{\hat{P}(Y_1 > Y_2)}{\hat{P}(Y_2 > Y_1)} = \frac{12(11 + 8 + 8) + 10(11 + 8) + 4(11)}{5(6 + 4 + 10) + 8(6 + 4) + 8(6)} = 2.45.$$

TABLE 2.2. Results of Study Comparing Two Treatments for Gastric Ulcer

Treatment Group	Change in Size of Ulcer Crater				
	Larger	$<\frac{2}{3}$ Healed	$\geq\frac{2}{3}$ Healed	Healed	Total
A	6	4	10	12	32
B	11	8	8	5	32
Total	17	12	18	17	64

Source: Armitage (1955), with permission of the Biometric Society.

The sample number of pairs in which treatment A gave the better response equals 2.45 times the sample number of pairs in which treatment B gave the better response.

2.2   ORDINAL ODDS RATIOS FOR CONTINGENCY TABLES

In the preceding section we introduced an ordinal odds ratio for comparing two groups on an ordinal response. In this section we present alternative odds ratios for two-way cross-classifications of ordinal variables. Instead of a single odds ratio summarizing the entire table, this alternative approach provides a set of odds ratios for a table which, together with the marginal distributions, fully specifies the joint distribution.

Let's first briefly review the odds ratio for 2×2 tables. Within row 1, the sample odds that the response is in column 1 instead of column 2 equals $p_{1|1}/p_{2|1}$. Within row 2, the odds equals $p_{1|2}/p_{2|2}$. Each odds is nonnegative, with value greater than 1.0 when response 1 is more likely than response 2. The ratio of these odds is the sample odds ratio,

$$\hat{\theta} = \frac{p_{1|1}/p_{2|1}}{p_{1|2}/p_{2|2}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

The proportion of subjects that made response 1 is larger in row 1 than in row 2 if $\hat{\theta} > 1$, whereas it is smaller in row 1 if $\hat{\theta} < 1$. In a corresponding population, the two conditional distributions are identical if and only if $\theta = 1.0$.

2.2.1   Local, Global, and Cumulative Odds Ratios

For $r \times c$ tables, odds ratios can use each pair of rows in combination with each pair of columns. For rows a and b and columns c and d, the odds ratio $n_{ac}n_{bd}/n_{bc}n_{ad}$ uses four cells falling in a rectangular pattern. All such odds ratios of this type are determined by a basic set of $(r-1)(c-1)$ odds ratios. One such basic set consists of the odds ratios

$$\hat{\theta}_{ij} = \frac{n_{ij}n_{rc}}{n_{rj}n_{ic}}, \qquad i = 1, \ldots, r-1, \quad j = 1, \ldots, c-1, \qquad (2.3)$$

which use the cell in the last row and the last column as a baseline. Each odds ratio is formed using the rectangular array of cells determined by rows i and r and columns j and c (see Figure 2.1). For ordinal variables the odds ratio $\hat{\theta}_{11}$ for the four corner cells, which describes association with the most extreme categories of each variable, is often of particular interest. It compares the odds of the highest instead of the lowest response at the highest and lowest levels of the other variable. When the two variables have a positive or negative trend, this is often the strongest of the odds ratios (farthest from 1.0).

The construction for forming a minimal set of odds ratios that determine the entire set is not unique. A natural basic set of $(r-1)(c-1)$ odds ratios for two
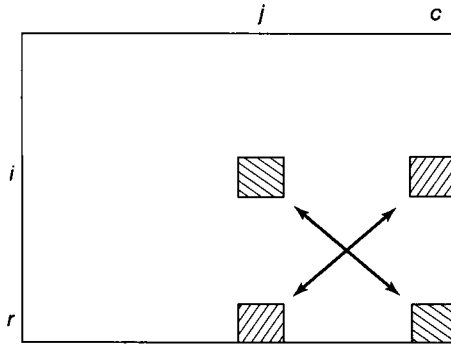
Figure 2.1. Odds ratios defined in (2.3).

ordinal variables is

$$\hat{\theta}_{ij}^{L} = \frac{n_{ij}n_{i+1,j+1}}{n_{i,j+1}n_{i+1,j}} \tag{2.4}$$

for $i = 1, \ldots, r - 1$, $j = 1, \ldots, c - 1$. These odds ratios use cells in adjacent rows and adjacent columns. Their values describe the relative magnitudes of associations in localized regions of the table. They are called *local odds ratios*.

A second natural family of odds ratios for ordinal variables is

$$\hat{\theta}_{ij}^{G} = \frac{\left(\sum_{a \leq i}\sum_{b \leq j} n_{ab}\right)\left(\sum_{a > i}\sum_{b > j} n_{ab}\right)}{\left(\sum_{a \leq i}\sum_{b > j} n_{ab}\right)\left(\sum_{a > i}\sum_{b \leq j} n_{ab}\right)}. \tag{2.5}$$

These measures are the regular odds ratios computed for the 2×2 tables obtained from the $(r - 1)(c - 1)$ ways of collapsing the row and column classifications into dichotomies. They describe associations that are global in both variables, in the sense that each odds ratio uses all categories of each variable instead of a localized region. They are called *global odds ratios*.

The local and global odds ratios treat row and column variables alike. They are especially useful when both variables are response variables. A family of odds ratios that distinguishes between rows and columns is

$$\hat{\theta}_{ij}^{C} = \frac{\left(\sum_{b \leq j} n_{ib}\right)\left(\sum_{b > j} n_{i+1,b}\right)}{\left(\sum_{b > j} n_{ib}\right)\left(\sum_{b \leq j} n_{i+1,b}\right)}. \tag{2.6}$$

These odds ratios are local in the row variable but global in the column variable. An equivalent definition for these odds ratios uses the sample conditional cumulative distribution functions of Y given x,

$$\hat{\theta}_{ij}^{C} = \frac{\hat{F}_{j|i}/(1 - \hat{F}_{j|i})}{\hat{F}_{j|i+1}/(1 - \hat{F}_{j|i+1})}. \tag{2.7}$$

We refer to them as *cumulative odds ratios*. These odds ratios are natural when x is an explanatory variable. They provide a comparison of pairs of levels of x with

respect to their entire conditional distribution on Y. For $2 \times c$ tables, global and cumulative odds ratios are identical.

Figure 2.2 illustrates local, global, and cumulative odds ratios. With positive counts, conversion of the cell counts into the set of odds ratios (2.3), (2.4),
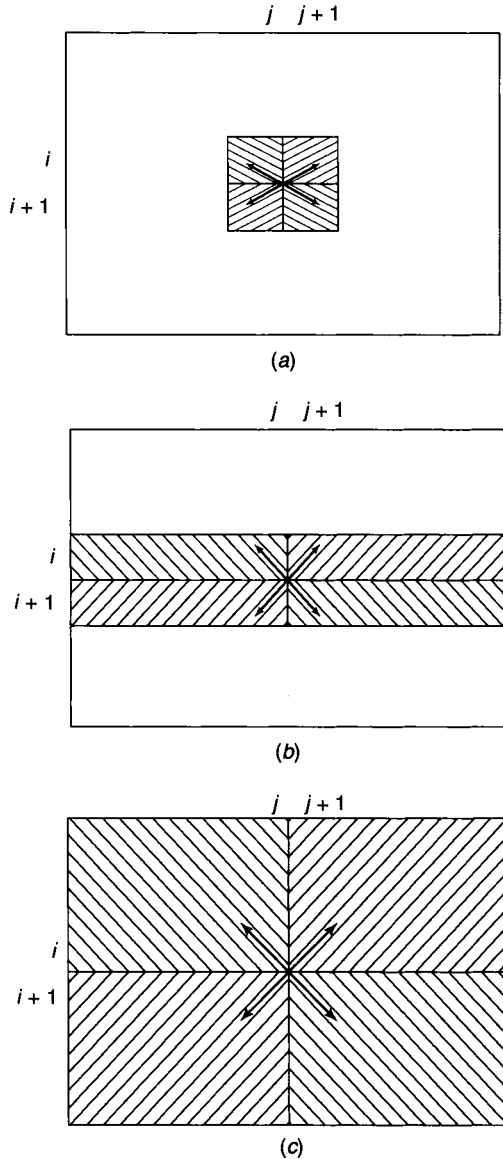
Figure 2.2. Three sets of $(r-1)(c-1)$ odds ratios for ordinal variables: (a) Local odds ratios, (b) cumulative odds ratios, and (c) global odds ratios.

(2.5), or (2.6) does not result in a loss of information. Given the marginal totals, the sample joint distribution of cell proportions or cell counts is determined by these odds ratios. For example, given sample global odds ratio values $\{\hat{\theta}_{ij}^G\}$ and sample marginal cumulative distribution functions $\hat{F}_i^X = p_{1+} + \cdots + p_{i+}$ and $\hat{F}_j^Y = p_{+1} + \cdots + p_{+j}$, the sample estimate of the joint distribution function $F_{ij} = P(X \le i, Y \le j)$ is

$$
\hat{F}_{ij} = \frac{1 + (\hat{\theta}_{ij}^G - 1)(\hat{F}_i^X + \hat{F}_j^Y) - \{[1 + (\hat{\theta}_{ij}^G - 1)(\hat{F}_i^X + \hat{F}_j^Y)]^2 - 4\hat{\theta}_{ij}^G(\hat{\theta}_{ij}^G - 1)\hat{F}_i^X \hat{F}_j^Y\}^{1/2}}{2(\hat{\theta}_{ij}^G - 1)}
$$

(2.8)

when $\hat{\theta}_{ij}^G \ne 1$ and $\hat{F}_{ij} = \hat{F}_i^X \hat{F}_j^Y$ when $\hat{\theta}_{ij}^G = 1$. The sample joint distribution determines the cell proportions.

2.2.2 Example: Happiness and Income

We illustrate odds ratios for ordinal variables with Table 2.3 from the 2006 General Social Survey. Respondents were asked, "Taken all together, would you say that you are very happy, pretty happy, or not too happy?" The table cross-classifies this response with family income, here measured as the response to the question, "Compared with American families in general, would you say that your family income is below average, average, or above average?"

Table 2.4 contains the sample values of the ordinal odds ratios. For example,

$$
\hat{\theta}_{11}^L = \frac{272 \times 835}{294 \times 454} = 1.70, \qquad \hat{\theta}_{11}^C = \frac{272 \times (835 + 131)}{(294 + 49) \times 454} = 1.69,
$$

$$
\hat{\theta}_{11}^G = \frac{272 \times (835 + 131 + 527 + 208)}{(294 + 49) \times (454 + 185)} = 2.11.
$$

These values mean that for those of above-average family income:

- The estimated odds of being very happy rather than pretty happy are $\hat{\theta}_{11}^L = 1.70$ times the corresponding estimated odds for those of average family income.

TABLE 2.3. Happiness and Relative Family Income

Family Income	Happiness			Total
	Very Happy	Pretty Happy	Not Too Happy	
Above average	272	294	49	615
Average	454	835	131	1420
Below average	185	527	208	920
Total	911	1656	388	2955

Source: 2006 General Social Survey.

TABLE 2.4. Values of Local, Cumulative, and Global Odds Ratios for Happiness Data of Table 2.3

	Local $\hat{\theta}_{ij}^L$		Cumulative $\hat{\theta}_{ij}^C$		Global $\hat{\theta}_{ij}^G$	
Row cut	$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$
$i = 1$	1.70	0.94	1.69	1.17	2.11	1.96
$i = 2$	1.55	2.52	1.87	2.87	2.20	3.01

- The estimated odds of being very happy rather than pretty happy or not happy are $\hat{\theta}_{11}^C = 1.69$ times the corresponding estimated odds for those of average family income.
- The estimated odds of being very happy rather than pretty happy or not happy are $\hat{\theta}_{11}^G = 2.11$ times the corresponding estimated odds for those of average or below-average family income.

All three sets of measures in Table 2.4 indicate that higher family income tends to be associated with higher happiness, except for $\hat{\theta}_{12}^L$, which is less than 1. This is also reflected by other summaries, such as the sample conditional distributions on happiness. For example, the estimated conditional probability of a "very happy" response takes value (0.44, 0.32, 0.20) for the family income levels (above average, average, below average).

2.2.3   Ordinal Odds Ratios Compare Numbers of Concordant and Discordant Pairs

Ordinal odds ratios provide various ways of dividing the number of concordant pairs of observations by the number of discordant pairs of observations. A pair of observations for two subjects is *concordant* if the subject ranking higher on X also ranks higher on Y. A pair of observations is *discordant* if the subject ranking higher on X ranks lower on Y. Each concordant pair gives evidence of a positive association, with *higher* values of X tending to occur with *higher* values of Y. Each discordant pair gives evidence of a negative association, with *higher* values of X tending to occur with *lower* values of Y.

For example, consider the global odds ratio, $\hat{\theta}_{11}^G$, for Table 2.3. For each variable, this particular global odds ratio treats the "high" and "low" dichotomy as "category 1" and "above category 1." The 272 observations in the cell that are "high" on both family income and happiness form concordant pairs when matched with each of the $(835 + 131 + 527 + 208)$ observations that are "low" on both variables. The $(294 + 49)$ observations that are "high" on family income but "low" on happiness form discordant pairs when matched with each of the $(454 + 185)$ observations that are "low" on family income but "high" on happiness. The total number of concordant pairs is $272 \times 1701 = 462{,}672$, the total number of discordant pairs is $343 \times 639 = 219{,}177$, and the first global odds ratio is $\hat{\theta}_{11}^G = 462{,}672/219{,}177 = 2.11$.

Each ordinal odds ratio has a particular identification of "high" and "low" for forming the concordant and discordant pairs. In Section 7.1 we show alternative ways of summarizing the two types of pairs that account simultaneously for all the possible ways of dichotomizing into "high" and "low."

2.2.4 Corresponding Population Ordinal Odds Ratios

With appropriate randomization in sampling or experiments, these sample ordinal odds ratios estimate corresponding odds ratios for a population. The population values can be defined in terms of joint probabilities $\{\pi_{ij}\}$ or conditional probabilities $\{\pi_{j|i}\}$. Joint probabilities are natural when both variables are response variables. Conditional probabilities are natural when one variable is explanatory.

The population local odds ratios are

$$\theta_{ij}^{L} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} = \frac{\pi_{j|i}/\pi_{j+1|i}}{\pi_{j|i+1}/\pi_{j+1|i+1}}. \tag{2.9}$$

Population cumulative odds ratios relate to the joint probabilities and to the conditional cumulative probabilities by

$$\theta_{ij}^{C} = \frac{\left(\sum_{b\le j}\pi_{ib}\right)\left(\sum_{b>j}\pi_{i+1,b}\right)}{\left(\sum_{b>j}\pi_{ib}\right)\left(\sum_{b\le j}\pi_{i+1,b}\right)} = \frac{F_{j|i}/(1-F_{j|i})}{F_{j|i+1}/(1-F_{j|i+1})}. \tag{2.10}$$

Because each global odds ratio uses all categories for each variable, this type of odds ratio makes sense for joint probabilities,

$$\theta_{ij}^{G} = \frac{\left(\sum_{a\le i}\sum_{b\le j}\pi_{ab}\right)\left(\sum_{a>i}\sum_{b>j}\pi_{ab}\right)}{\left(\sum_{a\le i}\sum_{b>j}\pi_{ab}\right)\left(\sum_{a>i}\sum_{b\le j}\pi_{ab}\right)}. \tag{2.11}$$

Let (X, Y) denote the row number and column number for an observation from the joint distribution $\{\pi_{ij}\}$. Then the global odds ratios are

$$\theta_{ij}^{G} = \frac{P(X \le i, Y \le j)P(X > i, Y > j)}{P(X \le i, Y > j)P(X > i, Y \le j)}.$$

By comparison, the local odds ratios are

$$\begin{aligned}\theta_{ij}^{L} &= \frac{P(X = i, Y = j)P(X = i+1, Y = j+1)}{P(X = i, Y = j+1)P(X = i+1, Y = j)} \\ &= \frac{P(Y = j+1 \mid X = i+1)/P(Y = j \mid X = i+1)}{P(Y = j+1 \mid X = i)/P(Y = j \mid X = i)},\end{aligned}$$

and the cumulative odds ratios are

$$\begin{aligned}\theta_{ij}^{C} &= \frac{P(X = i, Y \le j)P(X = i+1, Y > j)}{P(X = i, Y > j)P(X = i+1, Y \le j)} \\ &= \frac{P(Y \le j \mid X = i)/P(Y > j \mid X = i)}{P(Y \le j \mid X = i+1)/P(Y > j \mid X = i+1)}.\end{aligned}$$

Another ordinal odds ratio, less commonly used than the local, global, and cumulative odds ratios, is the *continuation odds ratio*,

$$\theta_{ij}^{CO} = \frac{P(Y = j \mid X = i)/P(Y > j \mid X = i)}{P(Y = j \mid X > i)/P(Y > j \mid X > i)}.$$

A separate and nonequivalent set of continuation odds ratios applies this formula after reversing the category order for both variables.

2.2.5   Stochastic Orderings of Groups

In comparing two groups, the notion of *stochastic ordering* is a way to characterize one group as being higher than the other on a quantitative response variable. The probability distribution for group 1 is *stochastically higher* than the probability distribution for group 2 if the cumulative distribution function (cdf) for group 1 is uniformly below the cdf for group 2. This means that for group 1, relatively more probability falls at the high end of the response scale. Figure 2.3 illustrates stochastic orderings for two groups with continuous probability density functions and cumulative distribution functions.

For an adjacent pair of rows i and $i + 1$ in a contingency table with ordinal response variable, the conditional distribution in row $i + 1$ is stochastically higher than the conditional distribution in row i if[3]

$$F_{j|i} \geq F_{j|i+1} \qquad \text{for } j = 1, 2, \ldots, c - 1.$$

Underlying
density functions

Underlying
distribution functions

(a)

(b)
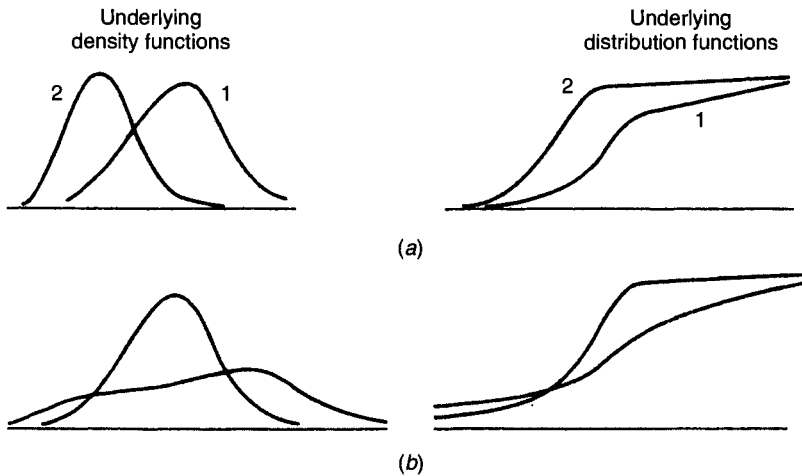
Figure 2.3. (a) Distribution 1 stochastically higher than distribution 2. (b) Distributions not stochastically ordered.

[3] At least one inequality should be strict, so the distributions are not identical.

This is equivalent to cumulative log odds ratios taking values

$$\log \theta_{ij}^C \geq 0 \quad (\text{hence } \theta_{ij}^C \geq 1) \qquad \text{for } j = 1, \ldots, c - 1.$$

Row $i + 1$ is *stochastically lower* than row i if $\log \theta_{ij}^C \leq 0$ for $j = 1, 2, \ldots, c - 1$. When the "high" end of the scale is column 1 instead of column c, the inequalities are reversed in these two definitions.

To illustrate, consider the sample cumulative odds ratios shown in Table 2.4. Treating the high end of the happiness scale as column 1, those of above-average family income are stochastically higher on happiness than those of average family income. Similarly, those of average family income are stochastically higher on happiness than those of below-average family income.

2.2.6 Types of Positive Dependence

For each type of odds ratio, X and Y with joint distribution $\{\pi_{ij}\}$ are *statistically independent* if all $(r - 1)(c - 1)$ of the population odds ratios equal 1.0, or equivalently, all population log odds ratios equal 0. Consider two ordinal variables that agree in terms of which end of the scale is regarded as the "high" end (i.e., either the first row and first column, or the last row and last column). For a given type of ordinal odds ratio, the association between the variables is called *positive* when all the log odds ratios are positive and *negative* when all the log odds ratios are negative.

Some definitions of positive (or negative) association are more stringent than others:

$$\text{If all } \log \hat{\theta}_{ij}^L > 0, \text{ then all } \log \hat{\theta}_{ij}^C > 0 \text{ and all } \log \hat{\theta}_{ij}^{CO} > 0.$$

$$\text{If all } \log \hat{\theta}_{ij}^C > 0 \text{ or if all } \log \hat{\theta}_{ij}^{CO} > 0, \text{ then all } \log \hat{\theta}_{ij}^G > 0.$$

For example, the set of tables having uniformly positive local log odds ratios is contained in the set of tables having uniformly positive cumulative log odds ratios, and that set is itself contained in the set of tables having uniformly positive global log odds ratios. The condition of positive association is therefore most stringent when expressed in terms of the local odds ratios. For a joint distribution in a $2 \times c$ case, when you reverse orientation of the table and consider the distribution of X given Y, the condition of uniformly positive local log odds ratios is equivalent to c strictly monotone-decreasing probabilities for category 1 of X.

When an association is positive (or negative) for all four ordinal odds ratios, more localized associations tend to be weaker in terms of log odds ratios being smaller in absolute value. For example, local log odds ratios tend to be weaker than cumulative log odds ratios, which tend to be weaker than global log odds ratios. Table 2.4 shows this behavior except for $\hat{\theta}_{11}^L = 1.70$ and $\hat{\theta}_{11}^C = 1.69$ being slightly out of order. For odds ratios [such as formula (2.3)] using just four cells in a rectangular pattern, the values tend to be stronger for less localized odds ratios.

For example, in Table 2.4 the local odds ratios fall between 0.94 and 2.52, whereas the sample odds ratio for the four corner cells in the table equals

$$\frac{n_{11}n_{33}}{n_{13}n_{31}} = \frac{272 \times 208}{185 \times 49} = 6.24.$$

See also Note 2.4.

When different researchers may choose different numbers of categories c for an ordinal response variable, an advantage of ordinal odds ratios that are global in Y is that their values do not usually depend much on that choice, since in each case they use the entire response scale. By contrast, each local odds ratio uses more of the response scale when c is small than when c is large. The local odds ratio is itself natural, though, when we want to make comparisons in terms of pairs of outcome categories rather than dichotomized regions of values.

2.3 CONFIDENCE INTERVALS FOR ORDINAL ASSOCIATION MEASURES

Next we present confidence intervals for ordinal measures. In Chapter 7 we consider this subject for a wider variety of ordinal measures of association. There we also present significance tests of independence of two ordinal variables for an alternative hypothesis expressed in terms of positive association for an ordinal log odds ratio. Tests and confidence intervals are also a by-product of various ordinal models presented in Chapters 3 and 4 that use these ordinal odds ratios to describe associations.

Let ζ denote a generic ordinal measure of association. For n observations its sample value, $\hat{\zeta}$, is a smooth function of sample proportions in the cells of a contingency table. We assume multinomial sampling over the cells of the contingency table; that is, the cell counts $\{n_{ij}\}$ have a multinomial distribution with parameters that are the cell probabilities $\{\pi_{ij}\}$. Then $\hat{\zeta}$ has an asymptotic (large-sample) normal sampling distribution by the *delta method* (Bishop et al. 1975, Sec. 14.6).

Let SE denote an estimated standard error for $\hat{\zeta}$. An approximate $100(1 - \alpha)\%$ confidence interval for ζ is

$$\hat{\zeta} \pm z_{\alpha/2}(\text{SE}),$$

where $z_{\alpha/2}$ denotes the standard normal percentile with right-tail probability equal to $\alpha/2$. For a 95% confidence interval, $\alpha = 0.05$ and $z_{0.025} = 1.96$.

Other confidence interval methods exist, discussed in Section 2.3.3, that are better in the sense that the actual coverage probability of the confidence interval tends to be closer to the nominal level. However, most software merely reports sample values of measures of association and their estimated standard errors, so this interval is simplest to obtain. The quality of the method often depends on the scale used, so this should be chosen with some care. For example, for odds ratios it is more sensible to use the asymptotic normality for the log odds ratio rather than

the odds ratio, because the sample log odds ratio converges much more quickly to normality as n increases.

2.3.1 Confidence Intervals for Ordinal Odds Ratios

In Section 2.2 we introduced four types of odds ratios for cross-classifications of ordinal variables: local odds ratios $\{\theta_{ij}^L\}$, global odds ratios $\{\theta_{ij}^G\}$, cumulative odds ratios $\{\theta_{ij}^C\}$, and continuation-ratio odds ratios $\{\theta_{ij}^{CO}\}$. For a particular ordinal odds ratio θ_{ij}, denote the four probabilities that make up the odds ratio by $\{\lambda_{ij}, \lambda_{i+1,j}, \lambda_{i,j+1}, \lambda_{i+1,j+1}\}$. Each ordinal odds ratio has the form

$$\frac{\lambda_{ij}\lambda_{i+1,j+1}}{\lambda_{i+1,j}\lambda_{i,j+1}}.$$

For example, for the local odds ratio θ_{ij}^L expressed in terms of joint probabilities,

$$\lambda_{ij} = \pi_{ij}, \quad \lambda_{i+1,j} = \pi_{i+1,j}, \quad \lambda_{i,j+1} = \pi_{i,j+1}, \quad \lambda_{i+1,j+1} = \pi_{i+1,j+1},$$

whereas for global odds ratio θ_{ij}^G,

$$\lambda_{ij} = \sum_{a \leq i}\sum_{b \leq j} \pi_{ab}, \quad \lambda_{i+1,j} = \sum_{a > i}\sum_{b \leq j} \pi_{ab},$$

$$\lambda_{i,j+1} = \sum_{a \leq i}\sum_{b > j} \pi_{ab}, \quad \lambda_{i+1,j+1} = \sum_{a > i}\sum_{b > j} \pi_{ab}.$$

Each type of ordinal odds ratio has the same form as the ordinary odds ratio for a 2×2 table. Ordinary inference for the odds ratio applies, with the probabilities used in the particular ordinal odds ratio. The estimated standard error for each ordinal log odds ratio is

$$\text{SE} = \sqrt{\frac{1}{n\hat{\lambda}_{ij}} + \frac{1}{n\hat{\lambda}_{i,j+1}} + \frac{1}{n\hat{\lambda}_{i+1,j}} + \frac{1}{n\hat{\lambda}_{i+1,j+1}}}. \tag{2.12}$$

When the region of the table covered by an odds ratio increases, the odds ratio has larger counts in the four cells. Thus, the sample log odds ratio value tends to be more precise as an estimator of the population value. For example, with standard sampling schemes, $\log\hat{\theta}_{ij}^G$ has smaller standard error than $\log\hat{\theta}_{ij}^C$ or $\log\hat{\theta}_{ij}^{CO}$, which have smaller standard errors than $\log\hat{\theta}_{ij}^L$.

A confidence interval for a log ordinal odds ratio is

$$\text{sample log odds ratio} \pm z_{\alpha/2}(\text{SE}). \tag{2.13}$$

Exponentiating (taking antilogs of) its endpoints provides a confidence interval for the odds ratio itself. When a sample odds ratio $\hat{\theta}$ equals 0 or ∞, the confidence

interval does not exist. When $\hat{\theta} = 0$, a sensible lower limit is 0. When $\hat{\theta} = \infty$, a sensible upper limit is ∞. The other bound can use the ordinary formula following some adjustment, such as replacing each $n\hat{\lambda}$ term that equals 0 in the SE formula by $\frac{1}{2}$. A less ad hoc approach forms the confidence interval by inverting score tests or likelihood-ratio tests about the value of the odds ratio, as discussed in Section 2.3.3. An alternative method uses a Bayesian approach, which naturally smooths the data based on prior beliefs and provides positive probability estimates in empty cells (Section 11.2).

2.3.2 Example: Happiness and Income Revisited

Table 2.3 showed a 3×3 table that cross-classified happiness with family income. Table 2.5 shows the four 2×2 tables that are collapsings of the original table formed to construct global odds ratios. Table 2.6 shows the sample global odds ratios (which were shown with other ordinal odds ratios in Table 2.4), the log global odds ratio values, the standard error estimates, and the 95% confidence intervals for the population global odds ratios. For example, for the first one ($\hat{\theta}_{11}^{G} = 2.11$),

$$SE = \sqrt{\frac{1}{272} + \frac{1}{343} + \frac{1}{639} + \frac{1}{1698}} = 0.094,$$

and the 95% confidence interval is

$$2.11[\exp \pm 1.96(0.094)] = (1.75, 2.53).$$

This set of confidence intervals suggests that there is a uniformly positive association between family income and happiness, as summarized by global odds ratios. But the association does not seem to be strong.

Here are two further considerations for these data: First, the values of the four sample global odds ratios are similar. One way to summarize the data further would be to form a weighted average of these values. This would be the outcome of fitting a model by which the four population global odds ratios are assumed to

TABLE 2.5. 2 × 2 Tables for Global Odds Ratios Between Happiness[a] and Relative Family Income[b]

Income[a]	Happiness[b]			
	VH	PH + NH	VH + PH	NH
AA	272	343	566	49
A + BA	639	1698	2001	339
AA + A	726	1309	1855	180
BA	185	735	712	208

[a]AA, above average; A, average; BA, below average.
[b]VH, very happy; PH, pretty happy; NH, not too happy.

TABLE 2.6. Global Odds Ratios for Table 2.5 and Corresponding Confidence Intervals

Odds Ratio	Sample Value	Log Odds Ratio	SE	95% CI
$\hat{\theta}_{11}^G$	2.11	0.745	0.094	(1.75, 2.53)
$\hat{\theta}_{12}^G$	1.96	0.671	0.160	(1.43, 2.68)
$\hat{\theta}_{21}^G$	2.20	0.790	0.094	(1.83, 2.65)
$\hat{\theta}_{22}^G$	3.01	1.102	0.111	(2.42, 3.74)

have identical values. We present such a model in Section 6.6. Second, if we truly want separate estimates such as those in Table 2.6, it can be useful to adjust the individual confidence levels so that the *overall* confidence level is controlled. A very simple way to do this uses a Bonferroni adjustment. If we want the overall confidence level to be at least 95% when we form confidence intervals for four global odds ratios, we would use error probability 0.05/4 for each individual interval (i.e., confidence level 98.75%).

2.3.3   Score and Profile Likelihood Confidence Intervals

The confidence interval of the form $\hat{\zeta} \pm z_{\alpha/2}(\text{SE})$ presented above is called a *Wald confidence interval*. It is based on inverting the *Wald test* of $H_0: \zeta = \zeta_0$ against $H_a: \zeta \neq \zeta_0$ using the large-sample normal test statistic

$$z = \frac{\hat{\zeta} - \zeta_0}{\text{SE}}.$$

For example, the 95% confidence interval consists of all ζ_0 values for which this test has a two-tailed P-value from the standard normal distribution that is larger than 0.05.

Wald confidence intervals for proportions or parameters based on proportions often perform poorly for small to moderate n. The actual coverage probability of a nominal 95% Wald confidence interval may be quite far from 0.95 unless n is quite large. This is especially true when ζ takes values near the boundary of the parameter space (such as in estimating a proportion that is near 0 or 1), in which case $\hat{\zeta}$ may have a highly skewed sampling distribution. Then it may not be sensible for $\hat{\zeta}$ to be the midpoint of the confidence interval, an extreme case being when $\hat{\zeta}$ falls at the boundary.

Alternative confidence intervals that provide results similar to those of Wald intervals for large n but usually perform better for small to moderate n result from inverting likelihood-ratio or score tests. The likelihood-ratio test has test statistic

$$-2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood values under the null hypothesis H_0 and under the alternative hypothesis H_a. The P-value is the right-tail

probability beyond the observed test statistic value, using the chi-squared distribution with df $= 1$. In this context, L_1 is the multinomial log-likelihood function evaluated at the sample proportions. The 95% *profile likelihood* confidence interval for ζ is the set of ζ_0 values for which the *P*-value > 0.05 for the likelihood-ratio test of H_0: $\zeta = \zeta_0$.

The score test is based on the derivative of the log-likelihood function and its standard error, evaluated at the null hypothesis value. The test statistic can often be expressed in the form

$$z = \frac{\hat{\zeta} - \zeta_0}{\text{SE}_0},$$

where SE_0 is the standard error estimated under the constraint that $\zeta = \zeta_0$. Although computationally more complex than the Wald method and not always readily available with software, these methods usually perform better in terms of having coverage probabilities closer to the nominal level. Lang (2008) provided a unified approach to fitting profile likelihood and score confidence intervals for contingency table parameters.[4]

We discuss this approach in the context of interval estimation of an odds ratio. Let θ denote a particular ordinal odds ratio, such as a global odds ratio for a 2×2 table of counts $\{m_{ij}\}$ that results from a particular collapsing of the $r \times c$ table $\{n_{ij}\}$. A 95% confidence interval based on inverting the likelihood-ratio (or score) test consists of all θ_0 values such that the *P*-value > 0.05 for the likelihood ratio (or score) test of H_0: $\theta = \theta_0$ against H_a: $\theta \neq \theta_0$. For each given θ_0 value there are expected frequencies $\{\hat{\mu}_{ij}(\theta_0)\}$ having the same margins as the observed 2×2 table $\{m_{ij}\}$ and having an odds ratio of θ_0. The likelihood-ratio statistic for H_0: $\theta = \theta_0$ has the form

$$G^2 = 2 \sum_i \sum_j m_{ij} \log \frac{m_{ij}}{\hat{\mu}_{ij}(\theta_0)}.$$

The 95% profile likelihood confidence interval consists of the set of θ_0 values for which G^2 is less than the 95th percentile of a chi-squared distribution with df $= 1$, which is 3.84 ($= 1.96^2$). The score test statistic for H_0: $\theta = \theta_0$ has the form of a Pearson chi-squared statistic,

$$X^2 = \sum_i \sum_j \frac{[m_{ij} - \hat{\mu}_{ij}(\theta_0)]^2}{\hat{\mu}_{ij}(\theta_0)}.$$

The 95% score confidence interval consists of the set of θ_0 values for which $X^2 \leq 3.84$. Simulations show that the score method works particularly well for estimating an odds ratio.

More generally, consider any hypothesis for a multinomial model that corresponds to a goodness-of-fit test. That is, the fitted values for the alternative

---

[4]Lang's R function *ci.table* can construct such confidence intervals.

hypothesis are the sample data, so testing the hypothesis corresponds to comparing fitted values under a null hypothesis to fitted values under an alternative hypothesis that are merely the sample cell counts. This is the case for the test about an odds ratio for a 2×2 table. Then the score statistic has the form of the Pearson statistic just shown (Smyth 2003; Lovison 2005).

Advantages of confidence intervals based on inverting likelihood-ratio tests or score tests is that unlike the Wald interval, they are not affected adversely when a sample odds ratio is 0 or ∞ and they do not depend on the scale. That is, applying them to the original scale or applying them to the log scale and then exponentiating yields the same result. Unfortunately, for some measures of association, standard statistical software does not yet provide score confidence intervals, but profile likelihood intervals are often available when the measure is equivalent to a parameter in a model. For the odds ratio, the profile likelihood confidence interval can be obtained with most logistic regression software by fitting a model with a binary predictor to the 2×2 table.[5] Although not readily available with some software, the score confidence interval for the odds ratio is relatively easy to obtain.[6]

2.3.4 Example: Comparing Treatments for Shoulder Pain

Table 2.7 comes from a study (Lumley 1996) to compare an active treatment with a control treatment for patients having shoulder tip pain after laparoscopic surgery. The two treatments were randomly assigned to 41 patients. The patients rated their pain level on the fifth day after the surgery.

Consider first the odds ratio for the odds that pain is in one of the first two categories instead of one of the last three. This is both a global odds ratio and a cumulative odds ratio. The sample odds ratio is 18.9, the estimated odds of a relatively low level of pain being much higher for the active treatment than for the placebo. The ordinary Wald confidence interval is (2.1, 170.4), the profile likelihood confidence interval is (3.0, 373.8), and the score confidence interval is (2.6, 128.1). With small samples, different methods can give quite different results. Based on simulations, we trust the score interval estimator of the odds ratio more than we do the other methods. With any of the intervals, we infer that the active treatment works better than the control treatment to reduce shoulder pain.

TABLE 2.7. Shoulder Tip Pain Scores After Laparoscopic Surgery

	Pain Score[a]				
Treatments	1	2	3	4	5
Active	19	2	1	0	0
Control	7	3	4	3	2

Source: Lumley (1996), Table 2.
[a] 1, low; 5, high.

[5] For example, in SAS, using the LRCI option in PROC GENMOD.
[6] An R function is available at www.stat.ufl.edu/~aa/cda/software.html.

2.3.5 Confidence Intervals for Measures Using $P(Y_1 > Y_2)$

We now consider the stochastic superiority measure $\alpha = P(Y_1 > Y_2) + \frac{1}{2} P(Y_1 = Y_2)$ for comparing two groups on an ordinal response. A confidence interval for α implies a corresponding confidence interval for $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$, since $\Delta = 2\alpha - 1$.

For independent multinomial samples of sizes n_1 and n_2 from the two rows, Halperin et al. (1989) showed that the sample value of the variance of the sample estimate $\hat{\alpha}$ is

$$
\mathrm{SE}^2 = \frac{1}{n_1 n_2}\left[\hat{\alpha} - (n_1 + n_2 - 1)\hat{\alpha}^2 + (n_2 - 1)C + (n_1 - 1)D - \frac{1}{4}\sum_{i=1}^{c} p_{i|1}\, p_{i|2}\right],
$$

$$(2.14)$$

where

$$
C = \sum_{i=1}^{c-1} p_{i|1}\left(\sum_{j=i+1}^{c} p_{j|2} + \frac{p_{i|2}}{2}\right)^2 + \frac{p_{c|1}\, p_{c|2}^2}{4},
$$

$$
D = \sum_{j=2}^{c} p_{j|2}\left(\sum_{i=1}^{j-1} p_{i|1} + \frac{p_{j|1}}{2}\right)^2 + \frac{p_{1|1}^2\, p_{1|2}}{4}.
$$

The Wald approach works better by applying it to estimate logit α rather than α. From the delta method, the 95% Wald confidence interval for logit α is

$$
\text{logit } \hat{\alpha} \pm 1.96 \frac{\mathrm{SE}}{\hat{\alpha}(1 - \hat{\alpha})}.
$$

Its bounds [LB, UB] induce the interval

$$
\left[\frac{\exp(\mathrm{LB})}{1 + \exp(\mathrm{LB})},\ \frac{\exp(\mathrm{UB})}{1 + \exp(\mathrm{UB})}\right]
$$

for α. If $\hat{\alpha}$ is either 0 or 1, the interval is undefined, and it is better to use the interval obtained with the profile likelihood or score method.

Ryu and Agresti (2008) used the result mentioned above about the score test statistic having the form of the Pearson statistic to construct a score confidence interval for α. For any given value α_0, the product multinomial likelihood can be maximized subject to the constraint $\alpha = \alpha_0$, leading to fitted values that can be compared to the observed counts with X^2. A 95% score confidence interval[7] is the set of α_0 values for which $X^2 \leq 3.84$.

For the shoulder pain data just analyzed, the 95% logit Wald confidence interval for α is (0.621, 0.874). The score confidence interval is (0.633, 0.875). The imprecision reflects the relatively small sample sizes.

---

[7]www.stat.ufl.edu/~aa/cda/software.html has R functions by E. Ryu for confidence intervals for α.

2.3.6 Small-Sample Interval Estimation for Local Odds Ratios

A well-known approach to small-sample inference for some parameters with categorical data eliminates unknown nuisance parameters by conditioning on their sufficient statistics. Statistical inference then uses the conditional distribution, which does not depend on the nuisance parameters. This method can be applied to interval estimation for odds ratios. With a multinomial distribution over the $r \times c$ table, conditioning on row and column totals yields a noncentral hypergeometric distribution that depends on the local odds ratios but not on unknown row and column marginal probabilities.

In Section 2.2.1 we mentioned that all odds ratios using a rectangular array of cells in a $r \times c$ table are determined by a basic set of $(r-1)(c-1)$ odds ratios, such as the odds ratios

$$\hat{\theta}_{ij} = \frac{n_{ij} n_{rc}}{n_{rj} n_{ic}}, \qquad i = 1, \ldots, r-1, \quad j = 1, \ldots, c-1.$$

With full multinomial sampling or independent multinomial sampling within rows or within columns, conditional on the marginal totals, the distribution of $\{n_{ij}\}$ is proportional to

$$\frac{\Pi_{i=1}^{r-1} \Pi_{j=1}^{c-1} \theta_{ij}^{n_{ij}}}{\Pi_{i=1}^{r} \Pi_{j=1}^{c} n_{ij}!}.$$

There is a one-to-one relationship between these odds ratios and the ordinal local odds ratios $\{\theta_{ij}^{L}\}$. The equivalent expression for this distribution in terms of the local odds ratios is

$$\frac{\Pi_{i=1}^{r-1} \Pi_{j=1}^{c-1} (\theta_{ij}^{L})^{s_{ij}}}{\Pi_{i=1}^{r} \Pi_{j=1}^{c} n_{ij}!},$$

where $s_{ij} = \sum_{a \leq i} \sum_{b \leq j} n_{ab}$.

A simple model considered in Section 6.2.2 assumes a common value $\beta = \log \theta_{ij}^{L}$ for the local log odds ratios. In that case, the conditional distribution of the data is proportional to $e^{\beta T} / \Pi_i \Pi_j n_{ij}!$, where $T = \sum_i \sum_j (ij) n_{ij}$. For fixed marginal totals, the maximum likelihood estimate $\hat{\beta}$ of β for this model is a strictly monotone function of T. This suggests basing exact inference for β on the conditional distribution of T. The statistic T is itself a monotone function of the correlation between X and Y for equally spaced scores for the rows and columns. As described in Section 7.6, we can base an exact conditional test on the conditional distribution of T. This can be done independent of any model for the odds ratios, but we consider a model of uniform local odds ratios here in order to consider a small-sample confidence interval for an ordinal effect measure.

Let C_t denote the sum of $(\Pi_i \Pi_j n_{ij}!)^{-1}$ for all tables with the given marginal totals that have $T = t$. Assuming a common local log odds ratio β, the conditional

distribution of T is (Agresti et al. 1990)

$$P(T = t \mid \{n_{i+}\}, \{n_{+j}\}; \beta) = \frac{C_t e^{\beta t}}{\sum_u C_u e^{\beta u}}.$$

To form a confidence interval for β, we invert exact tests of H_0: $\beta = \beta_0$ using this distribution. For observed value t_{obs} for T, an interval having confidence level at least $1 - \alpha$ is (β_ℓ, β_u), where β_ℓ and β_u are the solutions to the equations

$$\sum_{t \geq t_{obs}} P(T = t \mid \{n_{i+}\}, \{n_{+j}\}; \beta_\ell) = \frac{\alpha}{2} \quad \text{and}$$

$$\sum_{t \leq t_{obs}} P(T = t \mid \{n_{i+}\}, \{n_{+j}\}; \beta_u) = \frac{\alpha}{2},$$

except that $\beta_\ell = -\infty$ when T takes its minimum possible value (for the given margins) and $\beta_u = \infty$ when T takes its maximum possible value. The corresponding confidence interval for each local odds ratio, under the assumption that they are identical, is $(\exp(\beta_\ell), \exp(\beta_u))$. Mehta et al. (1992) presented a similar analysis for stratified $2 \times c$ tables with ordered columns.

When the sample size is very small or the data are unbalanced, with most observations in a single row or column, the inference can be quite conservative. The actual confidence level can be much larger than the nominal value. To achieve a confidence level that tends to be closer to the nominal level, although no longer guaranteed to be at least that level, invert the test using the mid-P value. That is, in the tail sums of probabilities just given, include only $\frac{1}{2} P(T = t_{obs} \mid \{n_{i+}\}, \{n_{+j}\}; \beta_u)$. The mid-$P$ interval also has the advantage of being a bit shorter.

For 2×2 tables, software can easily obtain small-sample confidence intervals for the odds ratio by conditioning on the marginal counts. Thus, it is possible to construct such confidence intervals for any particular ordinal odds ratio introduced in Section 2.2. Alternatively, an unconditional approach can be used for small-sample confidence intervals for odds ratios. Agresti (2002, pp. 99–100) has details and references. However, that method and the mid-P-based conditional method are not yet available in standard software.

2.3.7 Example: Severity of GVHD in Leukemia Patients

The StatXact manual (2005) reports Table 2.8, from one protocol for a study at the Dana Farber Cancer Institute. For patients receiving a bone marrow transplant, the ordinal response was the severity of graft versus host disease (GVHD). Table 2.8 covers a suspected risk factor: whether there was a type of blood incompatibility, called a *MHC mismatch*, between the donor and the recipient of the bone marrow.

The sample local odds ratios for the table, calculated as the estimated odds of the worse of two adjacent outcomes for mismatch divided by the corresponding estimated odds for match, are

$$\frac{2 \times 3}{2 \times 4} = 0.75, \quad \frac{2 \times 4}{2 \times 1} = 4.0, \quad \frac{1 \times 1}{2 \times 2} = 0.25, \quad \frac{1 \times 2}{1 \times 0} = \infty.$$

TABLE 2.8. Severity of GVHD in Leukemia Patients by Whether Patient Had MHC Mismatch

MHC Status	Severity of GVHD Toxicity					Total
	None	Mild	Moderate	Severe	Extreme	
Mismatch	2	2	2	1	2	8
Match	3	4	1	2	0	10
Total	5	6	3	3	2	18

Source: StatXact (2005, p. 633), with permission.

Given the very small cell counts, these estimates are extremely imprecise and can benefit from smoothing. Under the assumption that the true distributions have a common value for the four local odds ratios, StatXact reports[8] a small-sample 95% confidence interval for that common local odds ratio of (0.58, 3.27). With such a small sample, it is not possible to estimate that odds ratio very precisely.

2.4 CONDITIONAL ASSOCIATION IN THREE-WAY TABLES

Most applications have more than two relevant variables. The emphasis is often, then, on analyzing the relationship between the response variable and an explanatory variable at fixed levels of other explanatory variables or control variables. A *partial table* is a contingency table that displays counts for the relationship between two categorical variables at fixed levels of another variable (or variables). In a partial table, the other variable is controlled, in the sense that its value is held constant.

The association displayed in a partial table can be analyzed using the methods introduced in this chapter. For example, ordinal odds ratios apply to each partial table using the cell counts $\{n_{ijk}\}$ in the three-way contingency table that corresponds to "stacking up" the various partial tables. For the $(r-1)(c-1)$ local odds ratios at level k of the control variable(s), these are

$$\hat{\theta}^L_{ij|k} = \frac{n_{ijk}n_{i+1,j+1,k}}{n_{i,j+1,k}n_{i+1,j,k}}, \qquad i = 1,\ldots,r-1, \quad j = 1,\ldots,c-1. \qquad (2.15)$$

The response variable and explanatory variable are *conditionally independent*, given the control variable(s), if the population values of these odds ratios all equal 1.

2.4.1 Summary Measures of Conditional Association

If the association as described by an ordinal measure is similar in each partial table, it can be useful to pool the measure values into a summary measure of

---

[8]Using the "permutation with general scores" analysis and choosing scores 1, 2, 3, 4, and 5.

conditional association. This is also useful for meta-analyses, to combine results about an association from several studies.

One way to do this forms a weighted average of the sample values, with weights $\{w_k\}$ satisfying all $w_k > 0$ and $\sum_k w_k = 1$. For odds ratio measures, it is sensible to do this on the log scale. Let Z denote the control variable(s) (or variable identifying the various studies in a meta-analysis). Let K denote the number of categories of Z, which is also the number of partial tables. Possible choices for the weights $\{w_k\}$ include:

- $w_k = 1/K$, equal weight to the sample measure in each table.
- $w_k = n_{++k}/n$, the proportion of observations in the partial table.
- $w_k = (1/SE_k^2)/(\sum_{a=1}^{K} 1/SE_a^2)$, where SE_k is the estimated standard error of the sample measure in partial table k. The weight is inversely proportional to the estimated variance. This scheme approximates the measure in the class of weighted averages that has the smallest variance.
- For measures that are ratios, weights can be applied separately to the numerators and the denominators, such as is done in the Mantel–Haenszel odds ratio estimate for several 2×2 tables (e.g., Liu and Agresti 1996; Liu 2003).

Such summary measures have limited usefulness when there are multiple control variables. It is then more informative to use a modeling approach, as discussed starting in Chapter 3. This enables us to check the fit of the assumed association structure and to compare models of different complexities. For example, some models assume that the population ordinal odds ratios of a particular type in a two-way table are identical. If such a model fits each partial table well, we can analyze whether the extended model that has the same common ordinal odds ratio in each partial table fits well. If not, we would report the estimated odds ratios for the separate partial tables or use other ways of describing how the association varies across those tables.

2.4.2 Example: Association Between Political Views and Party, by Education

In 2006, the General Social Survey asked about the respondent's political ideology (liberal, moderate, conservative) and about the respondent's political party affiliation (Democrat, Independent, Republican). Table 2.9 summarizes the 4253 observations. Political party affiliation could be treated as nominal or ordinal. We treat it as ordinal to study whether there is a trend in ideology as one goes from Democrat to Republican. The sample conditional distributions, also shown in Table 2.9, suggest a moderately positive trend from liberal to conservative as one moves across the rows from Democrat to Republican. For a model presented in Section 6.2.2 that assumes a common value for all the local log odds ratios, the maximum likelihood estimate of that log odds ratio is 0.746. This corresponds to a local odds ratio estimate of $\exp(0.746) = 2.109$.

TABLE 2.9. Political Ideology by Political Party Identification, with Conditional Distributions on Political Ideology in Parentheses

Party	Political Ideology			Total
	Liberal	Moderate	Conservative	
Democrat	616 (44%)	522 (37%)	262 (19%)	1400
Independent	450 (26%)	821 (47%)	462 (27%)	1733
Republican	94 (8%)	305 (27%)	721 (64%)	1120

Source: 2006 General Social Survey.

TABLE 2.10. Values of Estimated Common Local Log Odds Ratio Between Political Ideology and Political Party Affiliation, Controlling for Education

Education	Sample Size	Local Log Odds Ratio
Less than high school	612	0.107
High school	2151	0.686
Junior college	362	0.754
Bachelor	734	1.200
Graduate	394	1.112
Overall	4253	0.746

Can this association be explained by education? For example, if more highly educated people tend to be more liberal and if more highly educated people tend to identify more as Democrats, perhaps this might explain the association and we may find little if any association at fixed levels of education.

Table 2.10 shows the estimated common local log odds ratios for the partial tables. The strength of association increases considerably across the education levels. There is little association for those with less than a high school education but a very strong association for those with a bachelor's or graduate degree. In this case, it seems better to report the separate values for the partial tables than to report a summary number for conditional association. Recall that less localized odds ratios would be stronger yet. For example, the estimated odds ratio using the four corner cells of Table 2.9 is $(616 \times 721)/(262 \times 94) = 18.0$. You can check that the local odds ratios relate to the corner odds ratio by $\hat{\theta}_{11}^{L} \hat{\theta}_{12}^{L} \hat{\theta}_{21}^{L} \hat{\theta}_{22}^{L}$. Thus, the model-based estimate of a common local odds ratio of 2.109 propagates to an estimated corner odds ratio of $2.109^4 = 19.8$, a very strong positive association.

2.5 CATEGORY CHOICE FOR ORDINAL VARIABLES

Most analyses presented in this book treat the ordinal scale as fixed, typically predetermined by the researcher who conducted the study. The results of some analyses may depend greatly, however, on the way the categories were defined.

In this section we illustrate this point, first for detecting the association between ordinal variables and then for describing a conditional association when a control variable is ordinal.

2.5.1 Finer Categorizations Provide More Power for Detecting Associations

With most ordinal variables, there are various possible scales for their measurement. Political ideology, for instance, might be measured with categories (liberal, moderate, conservative) or with categories (very liberal, liberal, slightly liberal, moderate, slightly conservative, conservative, very conservative). There are bias and power advantages to using categorizations having relatively more categories.

Often, it makes sense to imagine a continuous latent variable underlying the observed ordinal measurement. Then, an advantage of using more categories is that we get more information about the underlying effects. For example, as the numbers of rows and columns in a cross-classification of two ordinal variables are increased, the measurement gets finer. Then fewer pairs of observations are *tied*, falling in the same row or in the same column. (In Section 7.1.3 we present formulas for the various types of tied pairs.) Thus, more pairs of observations provide concordant or discordant indications that contribute to overall summaries such as ordinal odds ratio values.

Also, it is advantageous to have as many pairs untied as possible to increase power for determining the direction of the association. For example, an advantage of precise measurement of ordinal variables is that significance tests tend to be more powerful when there are relatively fewer tied pairs of observations. For testing the hypothesis of independence, the sample size needed to attain a certain power at a given significance level tends to decrease as r and c increase (Agresti 1976). This can be demonstrated by redefining categories for a table. We illustrate using Table 2.11, which is a 2×2 condensation of the 3×3 happiness data of Table 2.3. The "very happy" and "pretty happy" categories of happiness have been combined into the single category "happy." Also, the first and second income categories have been combined. Table 2.11 has log odds ratio equal to 1.102, with SE = 0.111. This is one of the global odds ratios of the original table. By comparison, for the original 3×3 table, a model considered in Section 6.6 that assumes a common value for all four global odds ratios has estimated log global odds ratios of 0.856, with SE = 0.068. With finer measurement, the ratio of the estimate to its SE is larger.

The values of ordinal odds ratios that are local for a variable tend to be more highly dependent on the categorization than are ordinal odds ratios that are global

TABLE 2.11. Condensation of Table 2.3 Used to Illustrate Effects of Category Choice

Family Income	Happiness	
	Happy	Not Too Happy
Average or above	1855	180
Below average	712	208

for that variable. Typically, the local odds ratios tend to be weaker when the response scale is finer.

Besides depending on the numbers of categories, the values of many measures of association depend on the marginal distributions of the variables, that is, on the relative numbers of observations in the different categories. This is the case for the odds ratios that group categories together. Because of this, it can be risky to compare values of measures calculated in tables having different category definitions or highly different marginal distributions. Consider, for example, case–control studies in which each subject who has a severe case of some disease is matched with someone having a mild case and a set of control subjects who do not have that disease, with all subjects observed in terms of some exposure that could cause the disease. The expected values for a summary measure of association would be different for a study that used one control for each pair of cases and a study that used more than one control for each pair of cases. An exception is the local odds ratio. It uses pairs of response categories and maintains the usual invariance to marginal proportions that is a well-known property of the odds ratio for 2×2 tables.

2.5.2 Finer Categorizations Describe Conditional Associations Better

Sometimes a control variable is also ordinal. When that variable represents categorical measurement of an underlying continuous variable, it is also advantageous to choose several categories for it to describe the conditional association adequately. For example, suppose that for an underlying continuous control variable Z, the value of a particular ordinal odds ratio between X and Y is identical at each fixed level of Z. We would want the measure value found with the ordinal categorization of Z to be relatively near the value that occurs for the underlying continuous measurement of Z. The approximation tends to improve as the number of categories of the control variable increases, since then Z is held more nearly constant in each partial table.

To illustrate, suppose that a trivariate normal distribution underlies three ordinal variables, with correlations $\rho_{XY} = 0.64$ and $\rho_{XZ} = \rho_{YZ} = 0.80$. For this distribution, the conditional distribution relating X and Y at a fixed level of Z is bivariate normal with partial correlation $\rho_{XY \cdot Z} = 0$. Then, for any way of categorizing X and Y at a fixed single value of Z, any ordinal odds ratio equals 1.0. Now, suppose that Z is not actually measured continuously but, rather, with K categories. Consider the value of the global odds ratio for the probabilities in the partial tables, when X and Y are dichotomized at the means of the underlying variables. Table 2.12 reports the global odds ratio values for those partial tables. These measures tend to approach 1.0 as the number of categories K of the control variable increases. However, with small K there can be substantial bias in approximating the underlying conditional association. We would probably fail to detect the absence of underlying conditional association if we used relatively few control categories or if one control category contained a majority of the observations. In this scenario we need $K \geq 5$ strata before the middle one has an odds ratio within 10% of the limiting value.

Even with relatively large K, Table 2.12 shows that a considerable conditional association can occur at the highest and lowest levels of the categorized control

TABLE 2.12. Global Odds Ratio for Partial Tables from a 2 × 2 × K Table Having an Underlying Trivariate Normal Distribution with Correlations 0.80 Between X and Z and Between Y and Z but Zero Partial Correlation Between X and Y

K	Marginal Probabilities for Z	Values of Global Odds Ratio
2	(0.1, 0.9)	1.66, 5.00
2	(0.5, 0.5)	2.20, 2.20
3	(1/3, 1/3, 1/3)	1.87, 1.29, 1.87
4	(0.25, 0.25, 0.25, 0.25)	1.77, 1.20, 1.20, 1.77
5	(0.2, 0.2, 0.2, 0.2, 0.2)	1.72, 1.15, 1.10, 1.15, 1.72
10	(0.1 each)	1.66, 1.11, 1.05, 1.03, 1.02, 1.02, 1.03, 1.05, 1.11, 1.66
∞		1.00 each

variable, when the marginal XY association is very strong. This tendency is not as severe when the marginal XY association is weaker. For example, suppose that $\rho_{XY} = 0.25$ and $\rho_{XZ} = \rho_{YZ} = 0.50$. Then again, $\rho_{XY \cdot Z} = 0$, and when $K = (2, 3, 4, 5, 10)$, the odds ratio in the partial table for the highest $1/K$ or lowest $1/K$ portion of the conditional distribution equals (1.28, 1.21, 1.19, 1.17, 1.13).

Cochran (1968) showed similar results for cases in which Y is quantitative and X is binary, in the context of reducing bias in comparing two groups in an observational study. When a quantitative variable can be measured in an essentially continuous manner, we are usually better off doing so rather than collapsing the variable into a few ordered categories. For models for ordinal response variables considered in this book it is possible to include continuous explanatory variables and control variables without having to categorize them.

2.5.3   Guidelines for Category Choice

Based on our experience, we suggest the following guidelines about category choice. These guidelines are approximate, and the exact behavior depends on the particular distributional structure. The guidelines are phrased in terms of measures of association, but they also apply to parameters describing associations in ordinal models presented in the remainder of the book.

- Ordinal categorical measures become relatively more efficient at detecting nonnull associations as r and c increase, since there are fewer tied pairs and standard errors tend to decrease.
- Measures of conditional association for ordinal categorical variables having ordinal control variables tend to become less biased in describing the conditional association for underlying continuous variables as more categories are used for the control variables.
- Different ordinal measures of association and ordinal models presented in the next four chapters typically yield similar conclusions about whether an

association exists when used in significance tests. However, the results of these analyses may depend strongly on how the categories are chosen for those ordinal variables.

CHAPTER NOTES

Section 2.1: Probabilities and Scores for an Ordered Categorical Scale

2.1. Other articles dealing with ridits include Brockett and Levine (1977), Vigderhous (1979), Semenya et al. (1983), Jansen (1984), Beder and Heim (1990), and Brunner and Puri (2001). For criticisms of the use of ridit scores, see Borgatta (1968), Mantel (1979), and Graubard and Korn (1987).

2.2. The stochastic superiority measure $\alpha = P(Y_1 > Y_2) + \frac{1}{2} P(Y_1 = Y_2)$ is popular in the nonparametric literature, both for comparing pairs of groups and for comparing each group to a marginal distribution. In the latter case it is the mean ridit, often called the *relative treatment effect*. It is used both for independent samples and repeated measures, as in Akritas and Brunner (1997), Brunner and Puri (2001), and Brunner et al. (2002). With $G_i(y) = [P(Y_i \leq y) + P(Y_i < y)]/2$, $i = 1, 2$, the *normalized distribution functions*, we can express

$$\alpha = \int G_2(y) dG_1(y),$$

an equation that also holds when Y_1 and Y_2 are continuous rather than ordinal categorical (Brunner and Munzel 2000). Simonoff et al. (1986) considered various estimators of Δ and their estimated variances, showing that a bootstrap method provides a robust estimated variance. For methods using such measures with several groups or multiple variables, see Semenya et al. (1983), Brunner and Puri (2001), Munzel and Hothorn (2001), Ryu and Agresti (2008), and several other articles by E. Brunner and colleagues summarized in Section 7.7.2. Bamber (1975) showed that α is the same as the area under a receiver operating characteristic (ROC) curve (Section 5.5.3).

2.3. Although α_{ik} from applying α to compare two groups i and k is not determined by the α values comparing group i to the marginal distribution of Y and comparing group k to the marginal distribution, models can be specified in which this type of simplicity occurs. An alternative to models considered in the next three chapters that use $r - 1$ parameters to compare r groups on an ordinal response variable [e.g., models (3.11) and (4.4)] is

$$\text{logit } (\alpha_{ik}) = \tau_i - \tau_k,$$

with a constraint such as $\tau_r = 0$. Semenya et al. (1983) proposed a weighted least squares analysis for this model. Kawaguchi and Koch (2010) generalized this model in the context of crossover studies.

Section 2.2: Ordinal Odds Ratios for Contingency Tables

2.4. In a $2 \times c$ table with an ordinal response Y, suppose that all $c - 1$ of the cumulative log odds ratios take value β. McCullagh and Nelder (1983, p. 122) noted that local log odds ratios $\{\log \theta_{1j}^{L}\}$ relate to the uniform cumulative log odds ratio β by

$$\log \theta_{1j} = \beta[P(Y \leq j + 1) - P(Y \leq j - 1)] + o(\beta), \qquad j = 1, \ldots, c - 1,$$

where $o(\beta)/\beta \to 0$ as $\beta \to 0$. Hence, local log odds ratios are typically smaller in absolute value than the cumulative log odds ratio β. For example, for the uniform marginal distribution $\{P(Y = j) = 1/c\}$, for small $|\beta|$, $\log \theta_{1j}^{L} \approx 2\beta/c$. The discrepancy between the two types of odds ratio tends to increase as c increases.

2.5. For other examples of ordinal odds ratios and for relationships among them, see Lehmann (1966), Dale (1984), Grove (1984, 1986), Douglas et al. (1990), Barnhart and Sampson (1994), and Oluyede (1994).

EXERCISES

2.1. Show that the mean ridits for conditional distributions of Y in a two-way contingency table satisfy $\sum p_{i+} \bar{A}_i = 0.50$.

2.2. For a sample set of counts $n_1, \ldots, n_c$, show that the jth midrank r_j relates to the jth ridit a_j by $r_j = na_j + 0.5$. Show that a_j and $r_j/(n + 1)$ are very close for large n.

2.3. Let Δ_{ik} denote the measure of stochastic superiority (2.2) applied to rows i and k of a contingency table.

(a) Does $\Delta_{ik} = \Delta_{ij} + \Delta_{jk}$? Why or why not?

(b) Show by example that it is possible to have $\Delta_{ij} > 0$ and $\Delta_{jk} > 0$, yet have $\Delta_{ik} < 0$.

2.4. For the local odds ratios, explain why $\theta_{ij} \geq 1$, $1 \leq j \leq c - 1$, implies that the conditional distribution in row $i + 1$ is stochastically higher than the conditional distribution in row i. Explain why the converse is not true.

2.5. For two conditional distributions, the plot of lines that connects successively the points $\{(0, 0), (F_{1|1}, F_{1|2}), (F_{2|1}, F_{2|2}), (F_{3|1}, F_{3|2}), \ldots, (1.0, 1.0)\}$ is called a *cumulative sum diagram* (CSD). See Grove (1980).

(a) Show that a straight line for the CSD corresponds to independence.

(b) Show that a convex CSD corresponds to the condition that all local log odds ratios for the two rows are nonnegative.

(c) Draw and interpret the sample CSD for the data in Table 2.2.

2.6. Lehmann (1966) defined two random variables (X, Y), discrete or continuous, to be *positively quadrant dependent* if

$$P(X \leq x, Y \leq y) \geq P(X \leq x)P(Y \leq y) \qquad \text{all } x \text{ and } y$$

and *positively likelihood-ratio dependent* if their joint density satisfies

$$f(x_1, y_1)f(x_2, y_2) \geq f(x_1, y_2)f(x_2, y_1)$$

whenever $x_1 < x_2$ and $y_1 < y_2$. He defined Y to be *positively regression dependent* on X if

$$P(Y \leq y \mid X = x) \qquad \text{is nonincreasing in } x.$$

(a) Show that the bivariate normal distribution satisfies positive likelihood-ratio dependence.

(b) For cross-classifications of ordinal variables, explain why positive quadrant dependence corresponds to $\{\log \hat{\theta}_{ij}^G \geq 0\}$, positive likelihood-ratio dependence corresponds to $\{\log \hat{\theta}_{ij}^L \geq 0\}$, and positive regression dependence corresponds to $\{\log \hat{\theta}_{ij}^C \geq 0\}$.

2.7. Go to sda.berkeley.edu/GSS and cross-classify the variables POLVIEWS and HAPPY for the latest survey [e.g., enter YEAR(2008) in the selection filter to obtain results for the year 2008]. Using methods of this chapter, describe the data.

CHAPTER 3

Logistic Regression Models Using Cumulative Logits

For binary response variables, in most fields logistic regression has become the standard model for analyzing the effects of explanatory variables. In Chapters 3 and 4 we present extensions of logistic regression for ordinal response variables. In Section 3.1 we describe ways of forming logits for an ordinal scale, in Sections 3.2 and 3.3 we present a model for one of these logits which applies to response cumulative probabilities, and in Section 3.4 we discuss model fitting and inference. In Section 3.5 we present model checking methods, in Section 3.6 introduce more complex models that sometimes fit better, and in Section 3.7 discuss connections between inference methods of models for cumulative probabilities and nonparametric rank methods. In Chapter 4 we present analogous models using the other ways of forming ordinal logits introduced in Section 3.1.

3.1 TYPES OF LOGITS FOR AN ORDINAL RESPONSE

When the response variable is ordinal, how can we form logits in a way that recognizes the category order? One way is to group categories that are contiguous on the ordinal scale. For example, we can apply the logit transformation to the cumulative probabilities.

3.1.1 Cumulative Logits

For c outcome categories with probabilities $\pi_1, \ldots, \pi_c$, the *cumulative logits* are defined as

$$
\begin{aligned}
\text{logit } [P(Y \leq j)] &= \log \frac{P(Y \leq j)}{1 - P(Y \leq j)} \\
&= \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}, \qquad j = 1, \ldots, c - 1.
\end{aligned} \tag{3.1}
$$

This logit equals the ordinary binary logit applied to the collapsing of the response outcomes into the two results $Y \leq j$ and $Y > j$. Each cumulative logit uses all c response categories.

3.1.2 Adjacent-Categories Logits

The *adjacent-categories logits* are the log odds for pairs of adjacent categories,

$$\log \frac{\pi_j}{\pi_{j+1}}, \qquad j = 1, \ldots, c - 1.$$

This logit equals the ordinary binary logit applied to the conditional probability of response outcome in category j, given response outcome in category j or $j + 1$; that is,

$$\text{logit} \left[P(Y = j \mid Y = j \text{ or } Y = j + 1) \right] = \log \frac{P(Y = j \mid Y = j \text{ or } Y = j + 1)}{1 - P(Y = j \mid Y = j \text{ or } Y = j + 1)}.$$

As a set of logits, the adjacent-categories logits are equivalent to the *baseline-category logits* commonly used to model nominal response variables. Those logits pair each category with a baseline category, typically the last one, as $\log(\pi_j/\pi_c)$, $j = 1, \ldots, c - 1$. The connections are

$$\log \frac{\pi_j}{\pi_c} = \log \frac{\pi_j}{\pi_{j+1}} + \log \frac{\pi_{j+1}}{\pi_{j+2}} + \cdots + \log \frac{\pi_{c-1}}{\pi_c} \qquad (3.2)$$

and

$$\log \frac{\pi_j}{\pi_{j+1}} = \log \frac{\pi_j}{\pi_c} - \log \frac{\pi_{j+1}}{\pi_c}, \qquad j = 1, \ldots, c - 1.$$

Either set is sufficient in the sense that it determines the logits for all $\binom{c}{2}$ pairs of response categories.

3.1.3 Continuation-Ratio Logits

The *continuation-ratio logits* are defined as

$$\log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_c}, \qquad j = 1, \ldots, c - 1. \qquad (3.3)$$

Continuation-ratio logit models are useful when a sequential mechanism determines the response outcome, in the sense that an observation must potentially occur in category j before it can occur in a higher category (Tutz 1991). An example is survival of a person through various age periods. Let $\omega_j = P(Y = j \mid Y \geq j)$. That is,

$$\omega_j = \frac{\pi_j}{\pi_j + \cdots + \pi_c}, \qquad j = 1, \ldots, c - 1. \qquad (3.4)$$

The continuation-ratio logits (3.3) are ordinary logits of these conditional probabilities: namely, $\log[\omega_j/(1 - \omega_j)]$.

An alternative set of continuation-ratio logits, appropriate if the sequential mechanism works in the reverse direction, is

$$\log \frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j}, \qquad j = 1, \ldots, c - 1. \tag{3.5}$$

The two forms of continuation-ratio logits are not equivalent. With $c = 3$ categories, for example, the first set (3.3) of sequential continuation-ratio logits is

$$\log \frac{\pi_1}{\pi_2 + \pi_3}, \qquad \log \frac{\pi_2}{\pi_3},$$

while the second set (3.5) is

$$\log \frac{\pi_2}{\pi_1}, \qquad \log \frac{\pi_3}{\pi_1 + \pi_2}.$$

3.1.4 Ordinal Models Use Ordinal Logits Simultaneously

For each type of ordinal logit applied to a c-category response, $c - 1$ logits can be formed. Ordinal models incorporate the $c - 1$ logits into a single model. In the next section we show that this approach results in more parsimonious and simpler-to-interpret models than the fitting of $c - 1$ separate models, one for each logit.

In this chapter we present models for cumulative logits and in Chapter 4 present models for the other ordinal logits. We see that each model has its own ordinal odds ratio for summarizing effects. For example, since the adjacent-categories logits use pairs of adjacent categories, they are naturally summarized using local odds ratios.

3.2 CUMULATIVE LOGIT MODELS

We now present a model for the cumulative logits, incorporating explanatory variables. For subject i, let y_i denote the outcome category for the response variable, and let $\mathbf{x}_i$ denote a column vector of the values of the explanatory variables. The model simultaneously uses all $c - 1$ cumulative logits. It has the form

$$\text{logit } [P(Y_i \leq j)] = \alpha_j + \boldsymbol{\beta}' \mathbf{x}_i = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \tag{3.6}$$

for $j = 1, \ldots, c - 1$, for a column vector $\boldsymbol{\beta}$ of parameters that describes the effects of the explanatory variables. For simplicity of notation, unless we need to refer to particular subjects or to particular values of the explanatory variables, we replace $P(Y_i \leq j \mid \mathbf{x}_i)$ in such equations by $P(Y \leq j)$, keeping in mind that in the model this is actually a conditional probability at each fixed value for the explanatory variables.

In model (3.6), the logit for cumulative probability j has its own intercept, α_j. The $\{\alpha_j\}$ are increasing in j because $P(Y \leq j)$ increases in j for each fixed value of $\mathbf{x}$, and the logit is an increasing function of this probability. The equivalent model expression for the cumulative probabilities is

$$P(Y \leq j) = \frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x})}, \qquad j = 1, \ldots, c - 1. \tag{3.7}$$

For the cell probabilities themselves,

$$P(Y = j) = \frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x})},$$

with $\alpha_0 = -\infty$ and $\alpha_c = \infty$. This formula has the form of a linear combination of inverse link functions: namely, inverse logit links with coefficients 1 and -1. The link function for the cell probabilities in such a case is called a *composite link function* (Thompson and Baker 1981).

In model (3.6), the effects $\boldsymbol{\beta}$ are the same for each cumulative logit. This results in a parsimonious model, compared to models such as baseline-category logit models for nominal responses that have separate parameters for each logit. We'll see motivation for this model structure in Section 3.3.2, based on an ordinary regression model for an underlying latent variable.

3.2.1   Cumulative Logit Model: Continuous Predictor

To help explain the cumulative logit model and its interpretations, let's first consider the case of a single continuous predictor x. The model is then

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \beta x, \qquad j = 1, \ldots, c - 1.$$

Figure 3.1 depicts this model for $c = 4$ outcome categories for Y. For fixed j, the response curve is an ordinary logistic regression curve for a binary response with outcomes $Y \leq j$ and $Y > j$. The common effect β for the three cumulative logits implies that the three response curves for the cumulative probabilities for $j = 1$, 2, 3 have the same shape.
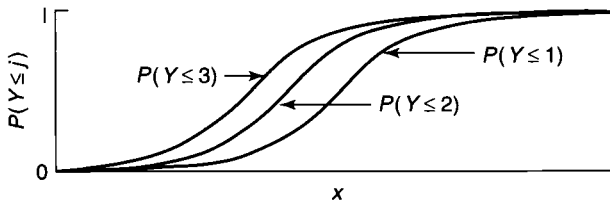
Figure 3.1. Depiction of cumulative probabilities in proportional odds version of cumulative logit model. Each curve has the same effect parameter.

As in logistic regression, the size of $|\beta|$ determines how quickly the curves go up or go down. At any fixed x value, the curves have the same ordering as the cumulative probabilities, the one for $P(Y \leq 1)$ being lowest. When the model holds with $\beta = 0$, the graph of $P(Y \leq j)$ as a function of x is a horizontal line for each j. Then Y is statistically independent of x.

Since the curves for the different cumulative probabilities have the same shape, any one curve is identical to any other curve shifted to the right or shifted to the left. For $j < k$, the curve for $P(Y \leq k)$ is the curve for $P(Y \leq j)$ translated by $(\alpha_k - \alpha_j)/\beta$ units in the x direction; that is,

$$P[Y \leq k \mid X = x] = P\left[Y \leq j \mid X = x + \frac{\alpha_k - \alpha_j}{\beta}\right].$$

The greater the difference $\alpha_k - \alpha_j$ for a given value of β, the greater the horizontal distance between the curves for the two cumulative probabilities. Although we need the intercept parameters $\{\alpha_j\}$ to fully determine the cumulative probabilities, in practice the parameter of interest is β, which describes the effect of x.

Figure 3.1 has $\beta > 0$ and Figure 3.2 shows corresponding curves for the category probabilities, $P(Y = j) = P(Y \leq j) - P(Y \leq j - 1)$. When $\beta < 0$, the analogous curves for Figure 3.1 descend rather than ascend, and the labels in Figure 3.2 reverse order. The identical shape for the various curves implies that the distributions of Y at different values of x are *stochastically ordered*, as defined in Section 2.2.5. If $\beta > 0$, each cumulative logit increases as x increases, which means that $P(Y \leq j)$
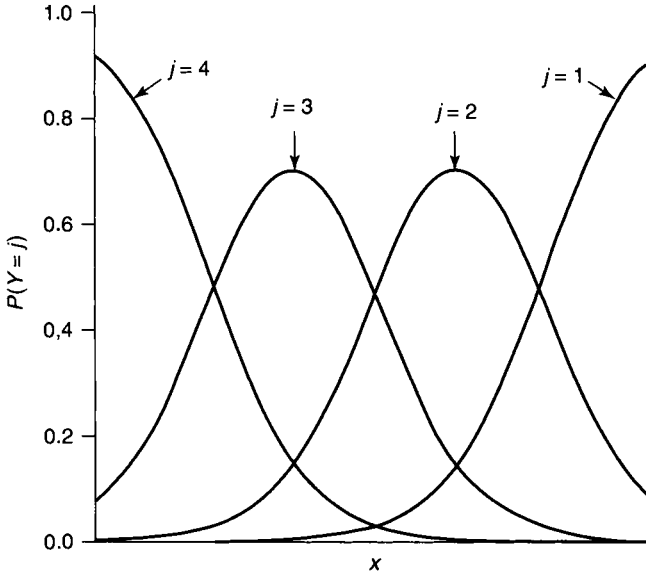
Figure 3.2. Depiction of category probabilities in proportional odds version of cumulative logit model. At any particular x-value, the four probabilities sum to 1.

increases also. This implies that the conditional Y distributions are stochastically lower at higher levels of x. If $\beta < 0$, the conditional Y distributions are stochastically higher at higher levels of x.

3.2.2 Alternative Parameterization with $-\beta'x$ Predictor

Often, the cumulative logit model is instead expressed as

$$\text{logit}\,[P(Y_i \leq j)] = \alpha_j - \beta'\mathbf{x}_i. \tag{3.8}$$

For this parameterization with a negative sign preceding $\beta_k x_{ik}$ for predictor k, the sign of β_k has the usual directional meaning. For example, $\beta_k > 0$ when Y is more likely to fall at the high end of the scale as x_{ik} increases. Specifically, if $\beta_k > 0$, then as x_{ik} increases, each cumulative logit decreases. Hence, the corresponding cumulative probabilities decrease. Then relatively less probability mass falls at the low end of the response scale, and Y is less likely to fall at the low end and more likely to fall at the high end of the scale.

Some software (such as SPSS) uses the linear predictor form $\alpha_j - \beta'\mathbf{x}_i$, whereas other software (such as SAS) uses $\alpha_j + \beta'\mathbf{x}_i$. Another way to have the usual sign interpretation for each effect is to express the model as[1]

$$\text{logit}\,[P(Y > j)] = \log \frac{P(Y > j)}{P(Y \leq j)} = \alpha_j + \beta'\mathbf{x}_i, \qquad j = 1, \ldots, c - 1,$$

with the cumulative probability in the denominator instead of the numerator.

3.2.3 Cumulative Logit Model for Contingency Table: Quantitative Predictor

Next we consider the cumulative logit model applied to a two-way contingency table. The rows are levels of a categorical explanatory variable. As in ordinary regression or logistic regression, if the explanatory variable is *qualitative* (nominal scale), indicator variables can represent its categories. The predictor is then a *factor* in the model. When the explanatory variable is *quantitative* with particular scores for the rows, such as the number of siblings of the subject, it is often sensible, instead, to represent that variable as a single predictor. Then a slope coefficient reflects a trend in Y as x changes. Similarly, although an ordinal variable can be treated as a factor, it is often useful to assign numerical scores to its categories and treat it in a quantitative manner. It is unnecessary to assign scores to the levels of Y, because the $c - 1$ cumulative logits within each row are ordered and serve as the responses.

---

[1]With the DESCENDING option, SAS fits the model in this form.

Consider first the case in which the explanatory variable is quantitative and we use ordered scores $\{u_1, \ldots, u_r\}$ for its rows. The cumulative logit model (3.6) then simplifies in terms of the row score u_i,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \beta u_i, \qquad j = 1, \ldots, c - 1. \tag{3.9}$$

As in ordinary logistic regression, interpretation of β can use log odds ratios. For rows a and b,

$$\text{logit } [P(Y \leq j \mid X = u_b)] - \text{logit } [P(Y \leq j \mid X = u_a)] = \beta(u_b - u_a).$$

This is the log odds ratio for the 2×2 table obtained using rows a and b and the binary response with outcomes $Y \leq j$ and $Y > j$. This log odds ratio is proportional to the distance between the rows, and for fixed a and b, it is the same for all j for collapsing the response.

With $a = i$ and $b = i + 1$, this log odds ratio is the *cumulative log odds ratio* estimated in equation (2.7). With unit-spaced row scores such as the row numbers $\{u_i = i\}$, the cumulative log odds ratio equals

$$\text{logit } [P(Y \leq j \mid X = u_{i+1})] - \text{logit } [P(Y \leq j \mid X = u_i)] = \beta \tag{3.10}$$

for $i = 1, \ldots, r - 1$, $j = 1, \ldots, c - 1$. Then $\exp(\beta)$ represents the constant value of the odds ratios $\{\theta_{ij}^C\}$ for the $(r - 1)(c - 1)$ separate 2×2 tables obtained by taking all pairs of adjacent rows and all binary collapsings of Y. These cumulative odds ratios take a uniform value whenever the row scores are equally spaced. We refer to this model for two-way contingency tables applied with equally spaced row scores as the cumulative logit *uniform association model*. Figure 3.3 illustrates the uniform cumulative odds ratio implied by this model.

3.2.4   Cumulative Logit Model for Contingency Table: Qualitative Predictor

When the explanatory variable is nominal scale, it enters the model as a factor, with indicator variables. The model has the form

$$\text{logit } [P(Y \leq j)] = \alpha_j + \tau_1 z_1 + \tau_2 z_2 + \cdots + \tau_{r-1} z_{r-1}, \qquad j = 1, \ldots, c - 1,$$
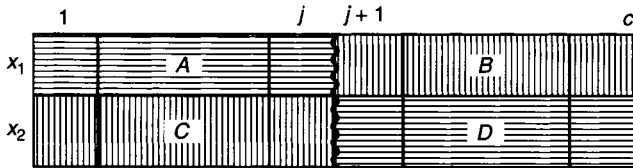
Figure 3.3. Odds ratios AD/BC that are constant for all pairs of adjacent rows and all $c - 1$ cumulative probabilities for a cumulative logit uniform association model.

in which $z_i = 1$ for an observation from row i and $z_i = 0$ otherwise. As usual, it would be redundant to include an indicator variable for the final category. The effect terms take the form of *row effects*. For short, we can express the model in terms of the effect for an observation in row i as

$$\text{logit } [P(Y \le j)] = \alpha_j + \tau_i \quad \text{for} \quad j = 1, \ldots, c - 1 \quad (3.11)$$

(Simon 1974). For identifiability, $\{\tau_i\}$ have a linear constraint, such as $\tau_r = 0$ or $\sum_i \tau_i = 0$. The constraint $\tau_r = 0$ corresponds to the model expressed above with indicator variables for each row except the last.

The row effect parameters determine the cumulative odds ratios. For a pair of rows a and b, the cumulative log odds ratio

$$\text{logit } [P(Y \le j \mid X = b)] - \text{logit } [P(Y \le j \mid X = a)] = \tau_b - \tau_a.$$

For these two rows, this cumulative log odds ratio is the same for all $c - 1$ of the 2×2 tables obtained for the $c - 1$ possible collapsings of the response to binary, $Y \le j$ and $Y > j$. If $\tau_b > \tau_a$, the cumulative probabilities are higher in row b than in row a, so the conditional Y distribution is stochastically lower in row b. Independence of Y and X is the special case $\tau_1 = \tau_2 = \cdots = \tau_r$.

3.2.5 Example: Astrology Beliefs and Educational Attainment

The 2006 General Social Survey asked subjects, "Would you say that astrology is very scientific, sort of scientific, or not at all scientific?" Table 3.1 cross-classifies their responses with their highest degree. The data show evidence of a trend, with more highly educated people tending to put less credence in astrology.

We treat opinion about astrology as the response variable. First we treat educational level as quantitative,

$$\text{logit } [P(Y \le j)] = \alpha_j + \beta u_i, \quad j = 1, 2, \quad (3.12)$$

TABLE 3.1. Education and Opinion About Astrology, with Conditional Distributions on Opinion in Parentheses

Highest Degree	Astrology Is Scientific		
	Very	Sort of	Not at All
< High school	23 (11%)	84 (41%)	98 (48%)
High school	50 (5%)	286 (31%)	574 (63%)
Junior college	4 (2%)	44 (26%)	122 (72%)
Bachelor	11 (3%)	57 (17%)	268 (80%)
Graduate	1 (1%)	23 (13%)	148 (86%)

Source: 2006 General Social Survey.

TABLE 3.2. Output for Fitting Cumulative Logit Models to Table 3.1

Parameter		DF	Estimate	Standard Error	Like. Ratio 95% Conf. Limits		Chi-Square
Intercept1		1	-2.3138	0.1246	-2.5636	-2.0746	344.77
Intercept2		1	-0.0264	0.0853	-0.1938	0.1408	0.10
degree		1	-0.4614	0.0486	-0.5581	-0.3674	90.09
Parameter		DF	Estimate	Standard Error	Like. Ratio 95% Conf. Limits		Chi-Square
Intercept1		1	-4.1236	0.2429	-4.6212	-3.6660	288.28
Intercept2		1	-1.8310	0.2196	-2.2860	-1.4212	69.50
degree	0	1	1.9439	0.2582	1.4524	2.4678	56.69
degree	1	1	1.2942	0.2299	0.8622	1.7673	31.68
degree	2	1	0.8782	0.2770	0.3443	1.4343	10.05
degree	3	1	0.4693	0.2581	-0.0241	0.9915	3.31
degree	4	0	0.0000	0.0000	0.0000	0.0000	.

using row scores (0, 1, 2, 3, 4), which are the values coded by the GSS. Table 3.2 shows software output for fitting models (as obtained with SAS, PROC GEN-MOD). The top part of the table shows results for this uniform association model (3.12). There are two intercept parameter estimates, with $\hat{\alpha}_1 = -2.3138$ and $\hat{\alpha}_2 = -0.0264$, because the response variable has three categories. The maximum likelihood estimate $\hat{\beta} = -0.4614$ reflects the tendency for the cumulative probability to decrease as the level of education increases. Each increase of a category in the level of attained education corresponds to a multiplicative impact of $e^{-0.4614} = 0.63$ in the estimated odds of response "very scientific" (instead of "sort of scientific" or "not at all scientific"), and in the estimated odds of response "very scientific" or "sort of scientific" (instead of "not at all scientific"). The estimated cumulative odds ratio comparing "< high school" with "graduate" education is $e^{(0-4)(-0.4614)} = 6.3$.

We can substitute the estimated parameters into the model formula to obtain estimated cumulative logits and then invert the estimated cumulative logits to obtain estimated cell probabilities. To illustrate, for "graduate" education, for which the predictor score equals 4,

$$\hat{P}(Y \leq 1) = \frac{e^{-2.3138+4(-0.4614)}}{1 + e^{-2.3138+4(-0.4614)}} = 0.015,$$

$$\hat{P}(Y \leq 2) = \frac{e^{-0.0264+4(-0.4614)}}{1 + e^{-0.0264+4(-0.4614)}} = 0.133.$$

From these,

$$\hat{P}(Y = 1) = 0.015, \quad \hat{P}(Y = 2) = 0.133 - 0.015 = 0.118,$$

$$\hat{P}(Y = 3) = 1 - 0.133 = 0.867.$$

The corresponding sample proportions are 0.006, 0.134, and 0.860.

Next, we treat educational level as qualitative (a factor) by fitting the model with row effects,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \tau_i, \qquad j = 1, 2.$$

The second panel of Table 3.2 shows results under the constraint $\tau_5 = 0$. The decrease in $\{\hat{\tau}_i\}$ across education categories again reflects the tendency for the cumulative probability to decrease as education level increases. The effect is strong for quite different educational levels. For example, the estimated cumulative odds ratio comparing "< high school" with "graduate" education is $e^{1.9439-0} = 7.0$.

3.3 PROPORTIONAL ODDS MODELS: PROPERTIES AND INTERPRETATIONS

We introduced cumulative logit models by focusing on simple models with a single explanatory variable. The property by which a cumulative log odds ratio comparing two settings of a predictor is identical for each of the $c - 1$ cumulative probabilities extends to the more general case (3.6), namely,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1.$$

3.3.1 Proportional Odds Property

The general model with multiple explanatory variables satisfies

$$\text{logit } [P(Y \leq j \mid \mathbf{x}_1)] - \text{logit } [P(Y \leq j \mid \mathbf{x}_2)]$$

$$= \log \frac{P(Y \leq j \mid \mathbf{x}_1)/P(Y > j \mid \mathbf{x}_1)}{P(Y \leq j \mid \mathbf{x}_2)/P(Y > j \mid \mathbf{x}_2)} = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2).$$

The odds of making response $Y \leq j$ at $\mathbf{x} = \mathbf{x}_1$ are $\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$ times the odds at $\mathbf{x} = \mathbf{x}_2$. The log cumulative odds ratio is proportional to the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$. The same proportionality constant applies to each of the $c - 1$ logits.

Because of this property, model (3.6) is often referred to as a *proportional odds model*. The name comes from an influential article on modeling ordinal data by McCullagh (1980). An alternative name used in some fields is *ordered logit model*. We prefer to refer to model (3.6) as the *proportional odds version of the cumulative logit model*. The term *ordered logit* is vague, because there are also other types of logit models for ordinal data, presented in Chapter 4. The term *proportional odds* is also vague, because these other logit models for ordinal data can also have a proportional odds structure.

3.3.2 Latent Variable Motivation

How can we justify the common effect β for different logits in the proportional odds version of the cumulative logit model? One way uses a regression model for

an unobserved continuous variable assumed to underlie Y (Anderson and Philips 1981). Let Y^* denote the underlying latent variable. For fixed values of explanatory variables $\mathbf{x}$, suppose that Y^* varies around a location parameter η, such as a mean, that depends on $\mathbf{x}$ through $\eta(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$. Specifically, suppose that the conditional cdf of Y^* is

$$P(Y^* \leq y^* \mid \mathbf{x}) = G(y^* - \eta) = G(y^* - \boldsymbol{\beta}'\mathbf{x}).$$

With the mean as the location parameter, at a given value of $\mathbf{x}$,

$$Y^* = \boldsymbol{\beta}'\mathbf{x} + \epsilon,$$

where ϵ has cdf G with $E(\epsilon) = 0$. Suppose that $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ are *cutpoints* of the continuous scale, sometimes also referred to as *thresholds*, such that the observed response Y satisfies

$$Y = j \qquad \text{if } \alpha_{j-1} < Y^* \leq \alpha_j.$$

That is, Y falls in category j when the latent variable falls in the jth interval of values. Figure 3.4 depicts this, and Figure 1.1 showed a particular example of a regression model for a latent variable and the corresponding ordinal categorical measurement.

Under this latent variable structure,

$$P(Y \leq j \mid \mathbf{x}) = P(Y^* \leq \alpha_j \mid \mathbf{x}) = G(\alpha_j - \boldsymbol{\beta}'\mathbf{x}).$$
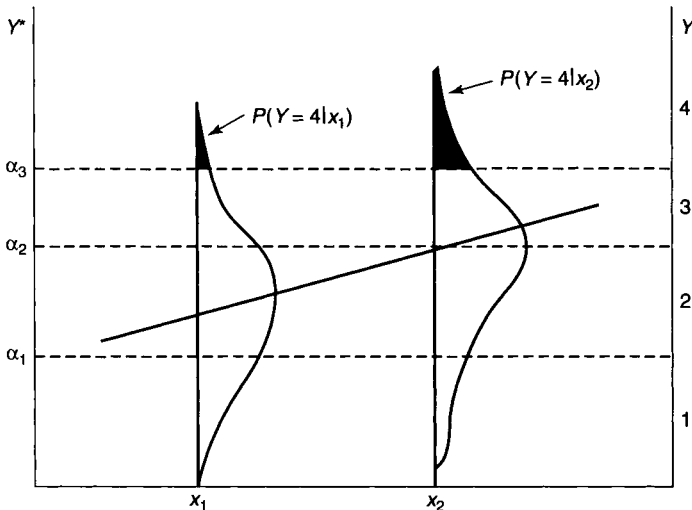

Figure 3.4. Ordinal measurement and underlying regression model for a latent variable.

So the link function to apply to $P(Y \leq j \mid \mathbf{x})$ to obtain a linear predictor is G^{-1}, the inverse of the cdf for Y^*. That is,

$$G^{-1}[P(Y \leq j \mid \mathbf{x})] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}.$$

When G is the cdf of the standard logistic distribution, which is $G(\epsilon) = e^\epsilon/(1 + e^\epsilon)$, then G^{-1} is the logit link function. That is, the logistic latent variable model implies the model for the observed response,

$$\text{logit } [P(Y \leq j \mid \mathbf{x})] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}.$$

This is the proportional odds version of the cumulative logit model, which has the same effects for each cumulative probability. Other underlying distributions imply different link functions for the cumulative probabilities but maintain the common effect for all j. In Section 5.1 we discuss the general model form. Most important, normality for ϵ implies that G^{-1} is the inverse of the standard normal cdf. This is the *probit link* for cumulative probabilities. In Section 5.2 we discuss cumulative probit models.

The derivation above shows that using a cdf of the form $G(y^* - \boldsymbol{\beta}'\mathbf{x})$ for the latent variable, with y^* values varying around $\boldsymbol{\beta}'\mathbf{x}$, results in linear predictor $\alpha_j - \boldsymbol{\beta}'\mathbf{x}$ rather than $\alpha_j + \boldsymbol{\beta}'\mathbf{x}$. In Section 3.2.2 we introduced this negative parameterization for the explanatory variables for the special case (3.8) with logit link. In practice, it does not matter which parameterization we adopt as long as we interpret effects appropriately.

The latent variable motivation for the model explains why distributions of Y at different settings of explanatory variables are stochastically ordered. The model is sensitive to location effects, not effects whereby the variability of Y changes as the explanatory variables change. The model usually fits poorly if the variability of an underlying latent variable changes dramatically over the range of observed values, as explained in Sections 3.6 and 5.4.

The latent variable construction also suggests an interpretation of the model parameters. A unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k, keeping fixed the other predictor values. The size of the effect depends on the spread of the conditional distribution of Y^* and can be specified in standard deviation units. When $\text{Var}(\epsilon) = \sigma^2$, a 1-unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k/σ standard deviations. The standard logistic distribution, for which the inverse gives the logit link function, has $\sigma = \pi/\sqrt{3}$. For the example in Section 3.2.5, an interpretation of $\hat{\beta} = -0.4614$ is that a 1-unit increase in educational level corresponds to an estimated decrease of $0.4614/(\pi/\sqrt{3}) = 0.25$ standard deviation for the mean of a hypothetical underlying latent conditional distribution for scientific belief in astrology.

McKelvey and Zavoina (1975) and Bock (1975, Sec. 8.1.6) suggested the latent variable representation for the normal case. Aitchison and Silvey (1957) and Ashford (1959) had used it with a single factor or quantitative variable as the predictor.

3.3.3 Invariance to Choice of Response Categories

In this derivation using a latent variable model, the same parameters β occur for the effects of the explanatory variables on Y regardless of how the $\{\alpha_j\}$ cut up the underlying continuous scale. This implies that the effect parameters β are invariant to the choice of categories for Y.

For example, suppose that a continuous variable measuring political ideology has a linear regression with some explanatory variables. Then the same effect parameters apply to a discrete version of political ideology with the categories (liberal, moderate, conservative) or (very liberal, slightly liberal, moderate, slightly conservative, very conservative). This feature makes it possible to compare model parameter estimates from studies using different response scales.

To illustrate, the cumulative logit uniform association model (3.12) for Table 3.1 has $\hat{\beta} = -0.461$. If we combine the categories "very scientific" and "sort of scientific" for opinion about astrology and fit the binary logistic regression model, we obtain $\hat{\beta} = -0.456$, very similar in value.

3.3.4 Interpretations Comparing Response Probabilities

An alternative way of summarizing effects in cumulative logit models directly uses the cumulative probabilities for Y. For example, to describe the effect of a quantitative variable x, we could compare $\hat{P}(Y \le j)$ [or $\hat{P}(Y > j)$] for a particular j at different values of x, such as the maximum and minimum values. To describe effects of a categorical predictor, we compare $\hat{P}(Y \le j)$ for different categories of that predictor. We can control for other quantitative variables in the model by setting them at their mean. We can control for other qualitative variables in the model by making the comparison at each combination of their values. When there are several qualitative variables, we could, instead, merely set them all at the means of their indicator variables, mimicking the treatment of quantitative control variables. Similarly, we can describe effects on the individual category probabilities. Using the lowest and highest categories of Y, we could report the maximum and minimum values of $\hat{P}(Y = 1)$ and $\hat{P}(Y = c)$ over the set of predictor values, reporting the values of the predictors that provide these extremes.

For example, consider the uniform association model (3.12) applied to Table 3.1. The estimated probability that astrology is judged to be "very scientific" decreases from 0.090 for those with less than a high school education to 0.015 for those with a graduate degree. The estimated probability that astrology is judged to be "not at all scientific" increases from 0.507 for those with less than a high school education to 0.867 for those with a graduate degree. Figure 3.5 portrays the estimated category probabilities on opinion about astrology at the five education levels. The height of the lowest portion (darkly shaded) of each bar is the estimated probability of response "not at all scientific"; the height of the medium bar is the cumulative probability for categories "not at all scientific" and "sort of scientific," so the middle portion of each bar portrays the estimated probability of response "sort of scientific." The more lightly shaded top portion of each bar portrays the estimated probability of response "very scientific." With multiple predictors, such
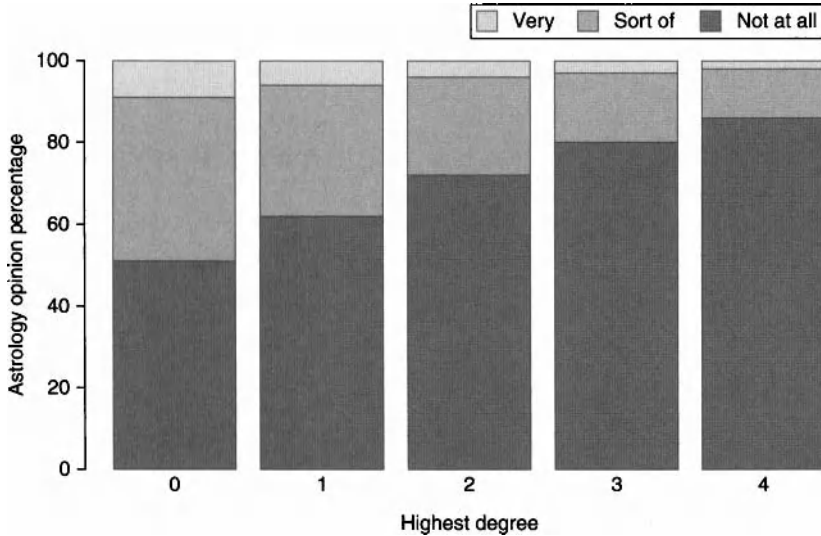
Figure 3.5. Estimated conditional probabilities for opinion about astrology at five highest degree levels based on fit of model (3.12). 0, less than high school; 1, high school; 2, junior college; 3, bachelor; 4, graduate.

a comparison can be made at the maximum and minimum values of each predictor, with the other predictors set at their means.

For a continuous predictor x_k, a comparison of probabilities at maximum and minimum values of x_k is not resistant to outliers on x_k. When severe outliers exist, it is often preferable to use the lower and upper quartiles of x_k instead. A comparison of estimated probabilities at the quartiles summarizes an effect over the middle half of the data (on x_k) and is not affected by outliers. Alternatively, a standard approximation for the rate of change of a probability in the logistic regression model also applies with ordinal logit models. The instantaneous rate of change in $P(Y \leq j)$ as a function of explanatory variable x_k, at fixed values for the other explanatory variables, is

$$\frac{\partial P(Y \leq j)}{\partial x_k} = \beta_k P(Y \leq j)[1 - P(Y \leq j)].$$

For example, suppose that $\hat{\beta}_k = 0.150$ for the effect of $x_k =$ number of years of education in a particular application. Then, at predictor values such that $\hat{P}(Y \leq j) = 0.60$, an increase of 1 in x_k while keeping fixed the other predictors corresponds to approximately a $0.150(0.60)(0.40) = 0.036$ estimated increase in $\hat{P}(Y \leq j)$.

We have seen that interpretations can also focus on standardized effects for the conditional distribution of an underlying latent response variable Y^*. Alternatively, standardized effects can refer to the marginal distribution of Y^*, as is often done in ordinary regression. This is discussed in Section 5.1.3. Yet another type of interpretation focuses on standardizing other measures. For example, Joffe and

Greenland (1995) showed how to convert estimated regression coefficients into estimates of standardized fitted probabilities, probability differences, and probability ratios.

3.4 FITTING AND INFERENCE FOR CUMULATIVE LOGIT MODELS

Next we discuss maximum likelihood (ML) fitting for cumulative logit models, assuming independent multinomial observations. The model of proportional odds form,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1,$$

constrains the $c - 1$ response curves to have the same shape. Because of this, its fit is not the same as the fits of $c - 1$ separate binary logistic models, one for each cumulative logit corresponding to a binary collapsing of the ordinal response.

Some early applications of cumulative logit models used *weighted least squares* for model fitting (e.g., Williams and Grizzle 1972). This entails applying the delta method to the sample proportions in the various categories at each setting of predictors to obtain a large-sample estimated covariance matrix of all the sample cumulative logits. Such logits are correlated with nonconstant variance, so ordinary least squares is not efficient. The weighted least squares approach has the advantage of computational simplicity, as the vector of model parameter estimates has closed form and does not require iterative methods. However, since the sample logits are functions of sample cell proportions, this approach is designed for nonsparse contingency tables and cannot handle continuous predictors. Walker and Duncan (1967) were the first to present ML model fitting for cumulative logit models. McCullagh (1980) presented an algorithm for more general models for cumulative probabilities discussed in Chapter 5.

3.4.1 Maximum Likelihood Model Fitting

For subject i, let $y_{i1}, \ldots, y_{ic}$ be binary indicators of the response, where $y_{ij} = 1$ for the category j in which the response falls; that is, when $Y_i = j$, then $y_{ij} = 1$ and $y_{ik} = 0$ for $k \neq j$. Recall that $\mathbf{x}_i$ denotes the values of the explanatory variables for subject i. Let $\pi_j(\mathbf{x}_i)$ denote $P(Y_i = j \mid \mathbf{X} = \mathbf{x}_i)$. For independent observations, the likelihood function is based on the product of the multinomial mass functions for the n subjects,

$$\prod_{i=1}^{n} \left[\prod_{j=1}^{c} \pi_j(\mathbf{x}_i)^{y_{ij}} \right] = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} \left[P(Y_i \leq j \mid \mathbf{x}_i) - P(Y_i \leq j - 1 \mid \mathbf{x}_i) \right]^{y_{ij}} \right\}$$

$$= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{c} \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} \right]^{y_{ij}} \right\}.$$

$$(3.13)$$

The likelihood function is a function of $(\{\alpha_j\}, \boldsymbol{\beta})$ after observing $\{y_{ij}\}$. Denote the log-likelihood function by $L(\{\alpha_j\}, \boldsymbol{\beta})$. We obtain each likelihood equation by differentiating L with respect to a particular parameter and equating the derivative to zero. For simplicity, we denote

$$G(z) = \frac{\exp(z)}{1 + \exp(z)}, \qquad g(z) = \frac{\exp(z)}{[1 + \exp(z)]^2}.$$

Then the likelihood equation for an effect parameter β_k is

$$\sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} x_{ik} \frac{g(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - g(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - G(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} = 0.$$

Section 5.1.2 shows the full set of equations in the context of a more general model with a family of link functions. As in ordinary logistic regression, these equations are nonlinear in the parameters and do not have a closed-form solution.

Iterative methods are used to solve the likelihood equations and obtain the ML estimates of the model parameters. Walker and Duncan (1967) and McCullagh (1980) used the *Fisher scoring algorithm* to do this. This is an iteratively reweighted least squares algorithm of the type used to fit ordinary generalized linear models. Each step has the form of weighted least squares, reflecting the nonconstant variance of the observations. The weights change from step to step as the approximations for the ML estimates of $\boldsymbol{\beta}$ get closer to the actual ML estimate $\hat{\boldsymbol{\beta}}$. Convergence of the iterative method is usually rapid because the log-likelihood function is concave (Haberman's discussion of McCullagh 1980; Burridge 1981; Pratt 1981; Kaufmann 1988).

McCullagh (1980) showed that sufficiently large n guarantees a unique maximum of the likelihood function. However, for finite n, unique estimates may not exist or may be infinite, for certain patterns of data, as explained in Section 3.4.5.

3.4.2 Estimating Standard Errors

The large-sample estimated covariance matrix for the ML model parameter estimates is the inverse of the *information matrix* evaluated at the ML estimates. The information matrix contains the negative second partial derivatives of $L(\{\alpha_j\}, \boldsymbol{\beta})$ with respect to the model parameters, that is, describing the curvature of the log-likelihood function. The more highly curved the log likelihood function at the ML estimates, the smaller are the standard errors, and the more precise are the ML estimates of the model parameters. In Section 5.1.2 we show this matrix in the context of a more general model with a family of link functions.

The information matrix has two possible versions. The *observed information matrix* uses the actual second partial derivatives. The element in row a and column b of the observed information matrix is $-\partial^2 L(\{\alpha_j\}, \boldsymbol{\beta})/\partial\beta_a \partial\beta_b$, where β_a and β_b are a pair of parameters from $(\alpha_1, \ldots, \alpha_{c-1}, \boldsymbol{\beta})$. By contrast, the *expected information matrix* uses the expected values of the second partial

derivatives. The element in row a and column b of the expected information matrix is $E[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_a \partial\beta_b]$. In either case the information matrix is estimated by substituting $(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})$. The estimated information matrix is inverted to obtain the estimated asymptotic covariance matrix. For either version of the inverse estimated information matrix, the estimated standard errors are the square roots of the main-diagonal entries.

The inverse of the expected information matrix is used in the Fisher scoring algorithm for obtaining the ML model fit. The corresponding algorithm that, instead, uses the observed information matrix is the *Newton–Raphson algorithm* [e.g., see Simon (1974) for the row effects model]. Results for $\hat{\boldsymbol{\beta}}$ are identical for each algorithm, as either algorithm yields the ML solution. For either algorithm, an estimated asymptotic covariance matrix is a by-product of the algorithm, from the inverse of the estimated information matrix at convergence. However, standard error estimates do depend on the algorithm used. For example, in SAS, PROC GENMOD uses the observed information, whereas PROC LOGISTIC uses the expected information, so their reported standard errors typically differ slightly. (By contrast, for binary logistic regression and baseline-category logit models, the observed information and the expected information are identical; see Agresti 2002, p. 149.) The Fisher scoring algorithm sometimes has better computational stability, because the weight matrix is positive definite over a larger region of the parameter space.

3.4.3 Inference About Model Parameters and Probabilities

Based on the model fit, we can conduct statistical inference about the model parameters using the ML estimates, their standard errors, and the maximized likelihood function in the usual ways. For example, a 95% Wald confidence interval for a parameter β_k is

$$\hat{\beta}_k \pm 1.96(\text{SE}),$$

where SE is the standard error of $\hat{\beta}_k$. For testing $H_0: \beta_k = 0$, we can use

$$z = \frac{\hat{\beta}_k}{\text{SE}}$$

or its square, which (under H_0) has an asymptotic chi-squared distribution with df $= 1$.

When the sample size is small or a large percentage of the observations fall at the highest (or lowest) category of the response variable, the distribution of $(\hat{\beta}_k - \beta_k)/\text{SE}$ need not be close to standard normal. Then it is better to use likelihood-ratio tests and confidence intervals based on the profile likelihood function. The likelihood-ratio test statistic equals

$$-2(L_0 - L_1),$$

where L_0 is the maximized log-likelihood function under the null hypothesis constraint that $\beta_k = 0$ and L_1 is the maximized log-likelihood function without that constraint, that is, evaluated at $\hat{\beta}$. The 95% confidence interval for β_k consists of null-hypothesis values β_{k0} for this parameter for which the P-value exceeds 0.05 for the likelihood-ratio test of H_0: $\beta_k = \beta_{k0}$. These inferences are available in some software.[2]

For the data on educational level and belief about astrology, Table 3.2 shows that the uniform association model for cumulative odds ratios has $\hat{\beta} = -0.4614$ with SE = 0.0486. The Wald 95% confidence interval for β is $-0.4614 \pm 1.96(0.0486)$, or $(-0.557, -0.366)$. The sample size was large, and this is similar to the profile likelihood confidence interval reported in that table. The corresponding confidence interval for the cumulative odds ratio for each one-category change in educational level is $(e^{-0.557}, e^{-0.366}) = (0.57, 0.69)$.

For interpreting effects, we can compare estimated cumulative probabilities or category probabilities at various settings of explanatory variables. Confidence intervals for the corresponding population probabilities describe the precision of those estimates. Some software also provides these inferences.[3]

3.4.4 Example: Mental Health by Life Events and SES

Table 3.3 comes from a study of mental health for a random sample of adult residents of Alachua County, Florida.[4] Mental impairment is ordinal, with categories (well, mild symptom formation, moderate symptom formation, impaired). The study related Y = mental impairment to several explanatory variables, two of which are shown here. The life events index x_1 is a composite measure of the number and severity of important life events that occurred to the subject within the past three years, such as the birth of a child, a new job, a divorce, or a death in the family. In this sample, x_1 has a mean of 4.3 and standard deviation of 2.7. Socioeconomic status (x_2 = SES) is measured here as binary.

The cumulative logit model of proportional odds form with main effects is

$$\text{logit } [P(Y \le j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2.$$

Table 3.4 shows SAS output. The estimates $\hat{\beta}_1 = -0.319$ and $\hat{\beta}_2 = 1.111$ suggest that the cumulative probability starting at the "well" end of the mental impairment scale decreases as life events increases and increases at the higher level of SES. Given the life events score, at the high SES level the estimated odds of mental impairment below any fixed level are $e^{1.111} = 3.0$ times the estimated odds at the low SES level.

Table 3.5 shows estimated category probabilities that also help us to interpret the predictor effects. To illustrate, we describe effects by the change in $\hat{P}(Y = 1)$

---

[2]For example, in SAS, using PROC GENMOD with the LRCI and TYPE3 options, and in R with the confint function.

[3]For example, in SAS, using PROC GENMOD with the OBSTATS option.

[4]Thanks to Charles Holzer for the background for this study.

TABLE 3.3. Mental Impairment by SES and Life Events

Subject	Mental Impairment	SES[a]	Life Events	Subject	Mental Impairment	SES[a]	Life Events
1	Well	1	1	21	Mild	1	9
2	Well	1	9	22	Mild	0	3
3	Well	1	4	23	Mild	1	3
4	Well	1	3	24	Mild	1	1
5	Well	0	2	25	Moderate	0	0
6	Well	1	0	26	Moderate	1	4
7	Well	0	1	27	Moderate	0	3
8	Well	1	3	28	Moderate	0	9
9	Well	1	3	29	Moderate	1	6
10	Well	1	7	30	Moderate	0	4
11	Well	0	1	31	Moderate	0	3
12	Well	0	2	32	Impaired	1	8
13	Mild	1	5	33	Impaired	1	2
14	Mild	0	6	34	Impaired	1	7
15	Mild	1	3	35	Impaired	0	5
16	Mild	0	1	36	Impaired	0	4
17	Mild	1	8	37	Impaired	0	4
18	Mild	1	2	38	Impaired	1	8
19	Mild	0	5	39	Impaired	0	8
20	Mild	1	5	40	Impaired	0	9

[a] 1, high; 0, low.

TABLE 3.4. Output for Fitting Cumulative Logit Model to Table 3.3

Parameter	Estimate	Std Error	Like. Ratio 95% Conf. Limits		Chi-Square	Pr > ChiSq
Intercept1	-0.2819	0.6423	-1.5615	0.9839	0.19	0.6607
Intercept2	1.2128	0.6607	-0.0507	2.5656	3.37	0.0664
Intercept3	2.2094	0.7210	0.8590	3.7123	9.39	0.0022
life	-0.3189	0.1210	-0.5718	-0.0920	6.95	0.0084
ses	1.1112	0.6109	-0.0641	2.3471	3.31	0.0689

```
                    LR Statistics
         Source      DF    Chi-Square    Pr > ChiSq
         life         1       7.78         0.0053
         ses          1       3.43         0.0641

      Score Test for the Proportional Odds Assumption
               Chi-Square       DF      Pr > ChiSq
                  2.3255         4        0.6761
```

TABLE 3.5. Estimated Probabilities Describing Effect of Life Events and SES on Mental Impairment

SES	Life Events	Estimated Probabilities			
		Well	Mild	Moderate	Impaired
High	Min. = 0	0.70	0.21	0.05	0.03
	Mean = 4.3	0.37	0.35	0.15	0.12
	Max. = 9	0.12	0.25	0.24	0.39
Low	Min. = 0	0.43	0.34	0.13	0.10
	Mean = 4.3	0.16	0.30	0.24	0.20
	Max. = 9	0.04	0.12	0.18	0.66

for the "well" outcome. Its value varies between 0.04, for those with low SES and nine life events, and 0.70, for those with high SES and zero life events. First, consider the SES effect. At the mean life events of 4.3, $\hat{P}(Y = 1) = 0.37$ at high SES (i.e., $x_2 = 1$) and $\hat{P}(Y = 1) = 0.16$ at low SES ($x_2 = 0$). Next, consider the life events effect. For high SES, $\hat{P}(Y = 1)$ changes from 0.70 to 0.12 between the sample minimum of zero and maximum of nine life events; for low SES, it changes from 0.43 to 0.04. Comparing 0.70 to 0.43 at the minimum life events and 0.12 to 0.04 at the maximum provides a further description of the SES effect. The sample effect is substantial for each predictor.

The precision of such estimates is portrayed by confidence intervals for population probabilities. Table 3.6 shows these for describing the life events effect on the probability of the "well" outcome. For the relatively small sample size of these data, the probability estimates are rather imprecise.

To illustrate inferential methods about effects of explanatory variables, we consider the effect of life events, controlling for SES. Table 3.4 reports a 95% profile likelihood confidence interval for β_1 of $(-0.572, -0.092)$. The confidence interval for the effect on the cumulative odds of a 1-unit increase in life events is $(\exp(-0.572), \exp(-0.092)) = (0.56, 0.91)$. The corresponding Wald confidence interval is $\exp[-0.3189 \pm 1.96(0.121)] = (0.57, 0.92)$, where the standard errors are based on the observed information (with model fitting using PROC GENMOD

TABLE 3.6. Estimates and Confidence Intervals Describing Effect of Life Events on Probability of "Well" Outcome

SES	Life Events	Estimated P(Well)	95% Confidence Interval
High	0	0.70	(0.39, 0.89)
	9	0.12	(0.03, 0.36)
Low	0	0.43	(0.18, 0.73)
	9	0.04	(0.01, 0.19)

in SAS). The chi-squared values reported in the table opposite the parameter estimates are for the Wald tests. For example, for testing H_0: $\beta_1 = 0$, the Wald statistic equals $(-0.319/0.121)^2 = 6.95$ with df $= 1$ (P-value $= 0.008$). The separate table for results of the likelihood-ratio tests also shows strong evidence of a life events effect but weaker evidence of an SES effect.

3.4.5 Infinite Model Parameter Estimates

In practice, with relatively small sample sizes, a large number of model parameters, or highly unbalanced data, one or more of the model parameter estimates may be infinite. An estimate $\hat{\beta}_k = \infty$ if the log-likelihood function continues to increase as β_k increases unboundedly, and $\hat{\beta}_k = -\infty$ if the log-likelihood function continues to increase as β_k decreases unboundedly. This happens most commonly with certain patterns of empty cells in contingency tables.

From binary logistic regression, we know that an estimate does not exist or is infinite when there is *quasi-separation*, that is, no overlap in the sets of explanatory variable values having $y = 0$ and having $y = 1$. A hyperplane passing through the space of predictor values can separate those with $y = 1$ from those with $y = 0$. For a cumulative logit model, this is the case if such separation occurs for each of the $c - 1$ collapsings of the ordinal response to a binary response.

For example, consider the quantitative predictor model (3.12), which is

$$\text{logit} [P(Y \leq j)] = \alpha_j + \beta u_i$$

with ordered scores $\{u_i\}$ for the rows. The estimate of β is infinite whenever either no pairs of observations are concordant or no pairs are discordant. One such case is a contingency table for which all observations fall in $r + c - 1$ cells consisting of one row and one column, with each at the highest or lowest level of the variable. Another case is when all observations fall on a diagonal of the table, such as an $r \times r$ table having $n_{ii} > 0$ for all i and $n_{ij} = 0$ for $i \neq j$.

Next, consider the qualitative predictor model (3.14) for an $r \times c$ table, which in terms of row effects $\{\tau_i\}$ is

$$\text{logit} [P(Y \leq j)] = \alpha_j + \tau_i$$

with $\tau_r = 0$. Infinite estimates exist if there is a pair of rows for which all observations in one row never fall above any observations in the other row. For example, $\hat{\tau}_i = \infty$ if all observations in row i fall in the first column and $\hat{\tau}_i = -\infty$ if all observations fall in the last column, when this does not happen for row r.

Whenever an infinite estimate exists for a given model, more complex models also have this property. For example, if $\hat{\beta}$ is infinite in the quantitative predictor model (3.12), finite effect estimates will not exist in the qualitative predictor model (3.11).

When infinite effect estimates occur, one solution is to use a simpler model for which all effect estimates are finite, such as by eliminating any predictor or

interaction term from the model that has an infinite estimate. Interpretations must then take this into account, and this is not a sensible solution if the simpler model fits poorly. Or, you can just use the model with an infinite estimate, realizing that this does not invalidate inference and prediction using methods other than Wald tests and confidence intervals. For example, suppose that β_k is a parameter for which $\hat{\beta}_k = \pm\infty$. Then the model still has a finite maximized log likelihood in the limit as $\hat{\beta}_k$ grows unboundedly, which software should report. So it is still possible to test H_0: $\beta_k = 0$ by comparing double the maximized log likelihood to its value for the simpler model. Similarly, it is still possible to obtain a profile likelihood confidence interval for β_k having form (L, ∞) when $\hat{\beta}_k = \infty$ and having form $(-\infty, U)$ when $\hat{\beta}_k = -\infty$. Finally, another solution is to fit the model with a Bayesian approach, which naturally shrinks model parameter estimates away from boundary values.

Most software does not recognize when $|\hat{\beta}_k| = \infty$. An indication of a likely infinite estimate is when a $|\hat{\beta}_k|$ reported is unusually large and the corresponding standard error is enormous. The iterative fitting process has then determined that it has reached the maximum of the log-likelihood function, and the relative flatness of the log-likelihood function at the point of convergence results in the extremely large SE value. If you are unsure, you can fit the corresponding binary logistic model for all possible binary collapsings of Y using software (such as PROC LOGISTIC in SAS) that can recognize when quasi-separation occurs and estimates are infinite.

3.4.6 Summarizing Predictive Power of Explanatory Variables

How can we summarize how well the response can be predicted using the fit of the model chosen? One way is with an index of predictive power called the *concordance index*. Consider all pairs of observations that have different response outcomes. The concordance index estimates the probability that the predictions and the outcomes are concordant, that is, that the observation with the larger y-value also has a stochastically higher set of estimated probabilities (and hence, for example, a higher mean for the estimated conditional distribution). The baseline value of 0.50 for the concordance index corresponds to its expected value from randomly guessing the response. A value of 1.0 results when knowing which observation in an untied pair has the stochastically higher estimated distribution enables us to predict perfectly which observation has the higher actual response. The higher the value of the concordance index, the better the predictive power.

Table 3.7 reports estimated concordance index values for some cumulative logit models fitted to the mental impairment data in Table 3.3. For the main effects model fitted above, for 70.5% of the untied pairs, the observation with the higher response outcome also had a stochastically higher estimated distribution. The predictive power was better than for models with a single predictor, but adding an interaction term provided no substantive improvement.

An alternative approach adapts standard measures for quantitative response variables, such as the multiple correlation and R-squared. For example, suppose that we assign scores $\{v_j\}$ to the outcome categories. Then we could find the correlation between the observed responses and the estimated means of the conditional

TABLE 3.7. Summary Measures of Predictive Power for Cumulative Logit Models Fitted to Mental Impairment Data of Table 3.3

Predictors in Model	Concordance Index	Multiple Correlation
SES	0.586	0.230
Life events	0.679	0.389
SES, life events	0.705	0.484
SES, life events, interaction	0.706	0.503

distributions from the model fit. This mimics the multiple correlation in multiple regression modeling. Or we could find the proportional reduction in variance when comparing the marginal variation to the conditional variation, which mimics R-squared (Agresti 1986). A related approach estimates R-squared for the regression model for an underlying latent response variable. McKelvey and Zavoina (1975) suggested this for the corresponding model with probit link function.

To illustrate, consider the no-interaction model with the mental impairment data in Table 3.3. The first subject in the sample has response in the first category ("well") and values 1 for life events and 1 for SES. From the prediction equation

$$\text{logit } [\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.319x_1 + 1.111x_2,$$

with $\hat{\alpha}_1 = -0.282$, $\hat{\alpha}_2 = 1.213$, and $\hat{\alpha}_3 = 2.209$, the estimated probabilities are (0.625, 0.256, 0.071, 0.047) for the four response categories. With scores (0, 1, 2, 3) for the categories (well, mild symptom formation, moderate symptom formation, impaired) of mental impairment, the observed response is 0 and the estimated mean response is $0(0.625) + 1(0.256) + 2(0.071) + 3(0.047) = 0.541$. For all 40 observations, the estimated correlation is 0.484 between the observed response and the predicted response given by the estimated conditional mean. Table 3.7 shows results with other models. Again, using both predictors provides improvement over a single predictor, but it does not help much to add an interaction term.

3.4.7 Classifying Observations into Ordered Categories

Some applications have values of several explanatory variables for a sample and require predictions for the category of an ordinal response that is not observed. For example, Marshall (1999) showed how to use explanatory variables such as body mass index, cholesterol levels, hypertension, and ethnicity to predict potential diabetes using the ordinal scale (normal, impaired glucose tolerance, diabetes). The standard test used to provide observations on this scale requires fasting and a blood test after a two-hour glucose load and is impractical for routine use as a screening instrument. Thus, a classification rule was considered useful for predicting the ordinal response based on the explanatory variables alone.

One approach to classification uses existing data for which the response is also observed to find an ordinal model that fits well, and then use the prediction equation

to generate estimated response probabilities. The classification rule would then predict the response category that has the highest estimated probability. In cases in which one outcome is much more likely than the others, this can result in always or nearly always predicting that category. Instead, Marshall (1999) classified using the category having the maximum estimated probability of the category divided by a prior probability for the category. He used this approach with the cumulative logit model as well as a classification tree approach and a search partition analysis method that was applied repeatedly to binary outcomes formed by collapsing adjacent categories of the ordinal scale. An evaluation of the methods indicated that the tree-based method had the largest overall misclassification rate and that classifications using the cumulative logit model performed well. See Note 3.6 for other literature on this topic.

3.5   CHECKING CUMULATIVE LOGIT MODELS

Having considered ML fitting and inference for cumulative logit models, we next present ways of checking the adequacy of the model fit. Methods include global goodness-of-fit tests as well as more narrowly directed methods such as model comparisons and residual analyses.

3.5.1   Testing Model Goodness of Fit for Contingency Tables

For nonsparse contingency tables, it is possible to conduct a goodness-of-fit test of the null hypothesis that the model holds against the alternative hypothesis that it does not. The alternative is equivalent to the saturated model, which fits the data perfectly. The test statistics compare the observed counts in the cells of the contingency table to expected frequency estimates based on the model fit.

At a particular setting $\mathbf{x}_i$ of the explanatory variables for which the observed multinomial sample has n_i observations, let $\{n_{ij}, j = 1, \ldots, c\}$ denote the observed cell counts for the c response categories. Under the null hypothesis that the model holds, the corresponding expected frequency estimates based on the model estimates of $\{P(Y = j \mid \mathbf{x}_i)\}$ equal

$$\hat{\mu}_{ij} = n_i \hat{P}(Y = j \mid \mathbf{x}_i), \qquad j = 1, \ldots, c.$$

The Pearson statistic for testing goodness of fit is

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

The corresponding likelihood-ratio (deviance) statistic is

$$G^2 = 2 \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}.$$

Under the null hypothesis that the model holds, X^2 and G^2 have large-sample chi-squared distributions. Their degrees of freedom equal the number of cumulative logits modeled minus the number of model parameters. The number of cumulative logits modeled equals the number of multinomial parameters in the saturated model: namely, $c - 1$ times the number of settings of the explanatory variables.

For example, a $r \times c$ contingency table has $c - 1$ multinomial parameters in each row, for a total of $r(c - 1)$ parameters. This is the number of parameters in the saturated model, for which the expected frequency estimates are merely the cell counts. The model (3.9) that treats the explanatory variable as quantitative,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \beta u_i,$$

has a single association parameter (β) and $c - 1$ intercept parameters (the $\{\alpha_j\}$) for the $c - 1$ logits, a total of c parameters. So the residual df for testing goodness of fit are df $= r(c - 1) - c = rc - r - c$. This is one less than the df $= (r - 1)(c - 1)$ for the independence model, which is the special case of this model with $\beta = 0$. Model (3.14), which treats the explanatory variable as qualitative with row effects,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \tau_i,$$

has $(c - 1) + (r - 1)$ parameters. Its residual df $= r(c - 1) - [(c - 1) + (r - 1)] = (r - 1)(c - 2)$, as noted by Simon (1974).

When the data are sparse or the model contains at least one continuous predictor, these global goodness-of-fit tests are not valid. Lipsitz et al. (1996) proposed an alternative goodness-of-fit test for such cases. It generalizes the Hosmer–Lemeshow test for binary logistic regression, which constructs a Pearson statistic comparing observed and fitted counts for a partition of such values according to the estimated probabilities of "success" using the original ungrouped data. This method does not seem to be available in current software. Pulkstenis and Robinson (2004) suggested an alternative approach. This is an area that still deserves serious research attention, to evaluate proposed methods and possibly develop others, such as normal approximations for chi-squared statistics when the data are very sparse.

3.5.2   Example: Astrology Beliefs and Education Revisited

For Table 3.1 on education and belief about astrology, in Section 3.2.5 we reported the fit of the uniform association model (3.12). For testing its goodness of fit, software (SAS, PROC LOGISTIC) reports

```
Criterion    Value    DF    Value/DF    Pr > ChiSq
  Deviance   5.8798    7     0.8400      0.5539
  Pearson    5.6081    7     0.8012      0.5862
```

The test statistic values $X^2 = 5.6$ and $G^2 = 5.9$ have df $= 7$ because the 10 multinomial parameters (two in each of the five rows of the table) are described by three parameters in the cumulative logit model (3.12). The model seems to fit well.

The more complex model (3.14) treats education as qualitative, using separate row effects. It has $X^2 = 3.7$ and $G^2 = 4.1$ with df $= 4$. The change in deviances compared to the simpler uniform association model, $5.9 - 4.1 = 1.8$ with df $= 7 - 4 = 3$, also indicates that the more complex model does not give a significantly better fit. The null hypothesis that the uniform association model holds is equivalent to stating that the qualitative predictor model holds with a linear trend in $\{\tau_i\}$.

3.5.3 Model Checking by Comparing Nested Models

If a model fails a goodness-of-fit test, various things could be wrong. Perhaps an important term is missing, such as an interaction term. Perhaps response distributions at different predictor values differ substantially in variability. Or perhaps, if the sample size is large, it is merely a matter of statistical significance without practical significance.

As with ordinary logistic regression, one way to check the model fit is to add terms and analyze whether the fit improves significantly. This approach is particularly useful when at least one explanatory variable is continuous or there are multiple predictors, because the goodness-of-fit tests presented above are invalid for data that are not contingency tables with reasonably sized counts. For a quantitative predictor, we could add a quadratic term or treat the predictor as a factor if it is categorical to allow for the effect to vary in a nonmonotone manner. For a multiple-predictor model, we could add interaction terms to the main effects. Comparing the working model to the more complex model can be done with a formal significance test. The likelihood-ratio test statistic equals $-2(L_0 - L_1)$, where L_0 is the maximized log likelihood under the simpler model. The df for the large-sample chi-squared distribution equals the number of extra parameters in the more complex model. Compared to conducting an overall goodness-of-fit test, an advantage of comparing the model to a more general model is that a small P-value suggests that the more general model be used as a new working model.

Alternatively, the comparison of models can use a criterion that summarizes how close the model's estimated cumulative probabilities are likely to fall to the true population values. The most popular such measure is AIC, the Akaike information criterion (see Section 3.5.9), which many software packages provide. As usual, compared to a more complex model, a more parsimonious model has benefits when the extra bias that the simpler model has is relatively small. The benefits include simplicity of description and possibly more precise estimation.

3.5.4 Example: Mental Health Modeling Revisited

In Section 3.4.4, using a cumulative logit model with main effects, we described how mental impairment (Y) depends on a quantitative life events index x_1 and a binary measurement of socioeconomic status (SES) x_2 ($1 =$ high, $0 =$ low). We can check the fit by comparing the model to more complex models. Permitting interaction yields a model with ML fit,

$$\text{logit } [\hat{P}(Y \le j)] = \hat{\alpha}_j - 0.420x_1 + 0.371x_2 + 0.181x_1x_2.$$

The coefficient 0.181 of $x_1 x_2$ has SE = 0.238. The estimated effect of life events is -0.420 for the low-SES group ($x_2 = 0$) and $(-0.420 + 0.181) = -0.239$ for the high-SES group ($x_2 = 1$). The impact of life events seems more severe for the low-SES group, but the sample size was relatively small and the estimates are imprecise. The likelihood-ratio statistic for testing H_0: $\beta_3 = 0$ for a lack of interaction is only 0.59 with df = 1 (P-value = 0.44). So the difference in effects is not significant, and the simpler model without an interaction term seems adequate.

3.5.5 Testing the Proportional Odds Assumption

Some software (such as PROC LOGISTIC in SAS) reports a score test of the proportional odds property (Peterson and Harrell 1990). This tests whether the effects are the same for each cumulative logit against the alternative of separate effects. It compares the proportional odds version of the model,

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1,$$

which has one parameter for each predictor, to the more complex model,

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}, \qquad j = 1, \ldots, c - 1. \tag{3.14}$$

With a single predictor, the proportional odds version of the model has one β, the more general model has $c - 1$ parameters $\{\beta_j\}$, and the large-sample chi-squared distribution for the score test has df $= (c - 1) - 1 = c - 2$. With p predictors, df $= p(c - 2)$.

The more complex model has the structural problem that cumulative probabilities can be out of order at some settings of the predictors. Because of this, it is often not feasible to maximize the likelihood function for model (3.14). Thus, the score test comparing the models is more widely applicable than a likelihood-ratio test or a Wald test, because the score test evaluates the rate of change of the log likelihood only at the null hypothesis, under which $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_{c-1}$. By contrast, the Wald and LR tests use the likelihood function maximized under the alternative hypothesis. When the more general model *can* be fitted, such tests can also be used, such as Wald tests proposed by Brant (1990) comparing separate estimates $\{\hat{\boldsymbol{\beta}}_j\}$ for each predictor.

To illustrate, Table 3.4 shows the result of this score test for the model fitted to the mental impairment data in Section 3.4.4 (using PROC LOGISTIC in SAS). This test compares the model with one parameter for x_1 and one for x_2 to the more complex model with three parameters for each, allowing different effects for logit $[P(Y \leq 1)]$, logit $[P(Y \leq 2)]$, and logit $[P(Y \leq 3)]$. The test statistic equals 2.33. It has df $= 4$, because the more complex model has four additional parameters. The more complex model does not fit significantly better ($P = 0.68$).

Unfortunately, this score test itself has limitations. First, Peterson and Harrell (1990) noted that the test may perform poorly for sparse data, such as when relatively few observations fall in one of the outcome categories or some explanatory

variables are continuous. Second, when the data are not sparse, its performance tends to be too liberal: P-values tend to be too small, and actual type I error rates tend to be greater than the nominal value.

Even if a model with nonhomogeneous effects for the different cumulative logits fits better over the observed range of x, for reasons of parsimony a simple model with proportional odds structure is sometimes preferable. One such case is when $\{\hat{\beta}_j\}$ for different logits with model (3.14) are not substantially different in practical terms. With a large sample size, a small P-value in the test of proportional odds may merely reflect statistical significance rather than practical significance. An analogy is with ordinary regression modeling using

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots .$$

This model often is useful for describing the linear component of the effect of a quantitative predictor, even though almost certainly the true relationship is more complex than linear. Similarly, the cumulative logit model with proportional odds structure often is effective in capturing the essence of the location effects even when the model has lack of fit, as illustrated by the example in Sections 3.5.6, 3.5.8, and 3.6.3.

For this reason, when the P-value is small in the test of proportional odds, it is useful to fit the model with nonhomogeneous effects or the ordinary binary logistic regression model for each of the $c - 1$ collapsings of the response. Compare the $c - 1$ estimated effects for each predictor to check whether some estimated effects vary greatly, such as changing direction in some cases. Then the more complex model may be useful. Alternatively, Kim (2003) suggested plotting estimated probabilities obtained under the proportional odds structure against the corresponding estimated probabilities found allowing different effects. In practical terms, the lack of fit is not severe if the pairs of estimated probabilities fall close to the line with intercept 0 and slope 1.

As explained in Section 3.6.1, biased effect estimators from the simple model may even have smaller mean-squared error than estimators from a more complex model, especially when the more complex model has a large number of additional parameters. So even if a test of proportional odds has a small P-value, we do not automatically reject the proportional odds form of the model.

3.5.6 Example: Religious Fundamentalism by Region

Table 3.8 cross-classifies subjects in the 2006 General Social Survey by the region in which they live and by whether they consider themselves fundamentalist, moderate, or liberal in their religious beliefs. Since region is of nominal scale type, we create indicator variables for its categories and consider the model (3.14) that treats it as a factor,

$$\text{logit} \, [P(Y \leq j)] = \alpha_j + \tau_i, \qquad j = 1, 2.$$

This model implies that the regions are stochastically ordered with respect to their distributions on religious beliefs. Table 3.8 also displays the sample conditional

TABLE 3.8. Data on Region of Residence and Religious Beliefs, with Conditional Distributions on Religious Beliefs in Parentheses

Region	Religious Beliefs		
	Fundamentalist	Moderate	Liberal
Northeast	92 (14%)	352 (52%)	234 (34%)
Midwest	274 (27%)	399 (40%)	326 (33%)
South	739 (44%)	536 (32%)	412 (24%)
West/Mountain	192 (20%)	423 (44%)	355 (37%)

Source: 2006 General Social Survey.

distributions. They show that each of the six pairs of regions are stochastically ordered except for (Northeast, West/Mountain). For that pair, in each extreme religious belief category (fundamentalist and liberal) the sample percentage is higher for West/Mountain than for Northeast.

Table 3.9 shows some SAS output from fitting the model. With the constraint $\hat{\tau}_4 = 0$, the ML estimates of row effects are $\hat{\tau}_1 = -0.07$, $\hat{\tau}_2 = 0.27$, $\hat{\tau}_3 = 0.89$. These show a tendency for fundamentalism to be much more common for subjects

TABLE 3.9. Output for Fitting Cumulative Logit Model of Proportional Odds Form to Table 3.8 on Residence and Religious Beliefs

```
                             Standard       Wald
Parameter       DF  Estimate    Error   Chi-Square   Pr > ChiSq
Intercept 1      1   -1.2618   0.0640   388.3248       <.0001
Intercept 2      1    0.4728   0.0611    59.8968       <.0001
region    1      1   -0.0702   0.0930     0.5687       0.4508
region    2      1    0.2688   0.0835    10.3531       0.0013
region    3      1    0.8896   0.0757   138.0873       <.0001

          Score Test for the Proportional Odds Assumption
                    Chi-Square       DF     Pr > ChiSq
                      93.0162         3       <.0001

       CELL-SPECIFIC STATISTICS
     Observed    Fitted    Stand. Resid.
y1        92    141.62       -7.788
y2       352    264.78        7.788
y3       234    271.60       -7.788
y4       274    270.04        0.567
y5       399    406.63       -0.567
y6       326    322.33        0.567
y7       739    688.34        7.902
y8       536    654.80       -7.902
y9       412    343.86        7.902
y10      192    214.04       -3.036
y11      423    383.54        3.036
y12      355    372.42       -3.036
```

in the South. The Northeast and West/Mountain states are similar but slightly less fundamentalist than the Midwest.

The score test of the proportional odds assumption compares this model with the more complex model having separate $\{\hat{\tau}_i\}$ for the two logits, that is, three extra parameters. From Table 3.9, the score test statistic equals 93.0 (df = 3), giving extremely strong evidence of lack of fit. The model with separate $\{\hat{\tau}_i\}$ for the two logits is saturated, so this test is an alternative to the Pearson and deviance statistics (not shown in the output table) for testing the model goodness of fit. Those statistics equal $X^2 = 97.5$ and $G^2 = 98.0$ (df = 3). In Sections 3.5.8 and 3.6.3 we investigate the nature of the lack of fit and whether it is substantively important.

3.5.7 Residuals to Detect Specific Lack of Fit

Global goodness-of-fit tests such as provided by the deviance have disadvantages. First, they do not apply with sparse contingency tables or when any explanatory variables are continuous. Second, even when the P-value is small, the test result gives no information about what's wrong with the model. The test of the proportional odds assumption that compares the model with the more general model replacing β by β_j is more directed than the global goodness-of-fit test when the more general model is not saturated.

An alternative way of checking for more specific types of lack of fit is to form residuals. For a contingency table with cell count n_{ij} and fitted value $\hat{\mu}_{ij}$ at setting i of the explanatory variables and response category j, the *standardized residual* is

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\text{SE}}.$$

The SE term is the estimated standard error of $n_{ij} - \hat{\mu}_{ij}$ under the presumption that the model holds. Lang (1996) provided an expression for SE. It uses an analog of the "hat" matrix for a class of generalized loglinear models, presented in Section 6.6.4, that contains cumulative logit models as a special case.

When the model holds, standardized residuals have approximate standard normal distributions. So relatively large values, such as exceeding about 3 in absolute value, indicate lack of fit in that cell. McCullagh (1980) suggested first finding the contribution provided to the deviance for each multinomial sample and then inspecting cell-specific residuals for those cases that have large contributions. In some cases, this may indicate that a model fits the data well except for one or two combinations of explanatory variable values.

It can also be informative to form residuals using cumulative totals rather than cell counts. A standardized residual for a cumulative total at response category j for setting i of the explanatory variables divides $(\sum_{k=1}^{j} n_{ik} - \sum_{k=1}^{j} \hat{\mu}_{ik})$ by its SE. More simply, a Pearson-type residual has the form

$$\frac{\sum_{k=1}^{j} n_{ik} - \sum_{k=1}^{j} \hat{\mu}_{ik}}{\sqrt{n_i \hat{P}(Y \le j \mid \mathbf{x}_i)[1 - \hat{P}(Y \le j \mid \mathbf{x}_i)]}} = \frac{\sum_{k=1}^{j} n_{ik} - n_i \hat{P}(Y \le j \mid \mathbf{x}_i)}{\sqrt{n_i \hat{P}(Y \le j \mid \mathbf{x}_i)[1 - \hat{P}(Y \le j \mid \mathbf{x}_i)]}},$$

where the denominator is the estimated standard deviation of $\sum_{k=1}^{j} n_{ik}$ based on the model fit. However, this residual does not have the standard normal as its reference distribution because it does not account for $\sum_{k=1}^{j} \mu_{ik}$ being estimated in the numerator. Or, we could informally inspect cumulative sums of the cell-specific standardized residuals defined above. Liu et al. (2009) proposed graphical diagnostics based on cumulative sums of residuals to diagnose misspecification for the proportional odds version of the cumulative logit model. Pruscha (1994) proposed partial residuals. Bender and Benner (2000) proposed various graphics, including smoothed partial residual plots.

3.5.8 Example: Religious Fundamentalism by Region Revisited

Let's now investigate the nature of the lack of fit of the row effects model to Table 3.8. We first consider the contribution to the deviance of each row. In a given row i, this contribution

$$2 \sum_j n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}$$

equals 51.4 in row 1, 0.2 in row 2, 39.3 in row 3, and 7.1 in row 4. Rows 1 and 3 both contribute strongly to the overall lack of fit.

Next we inspect the cellwise standardized residuals for the fit of the model, $r_{ij} = (n_{ij} - \hat{\mu}_{ij})/\text{SE}$, focusing on rows 1 and 3. Table 3.9 also shows the fitted values and the standardized residuals.[5] With residual df $= 3$, there are many redundancies among their values, and each row has only one bit of information about lack of fit. From these, we clearly see the nature of the lack of fit. The Northeast (and West/Mountain states) has more people in the moderate category and fewer in the other two categories than the model predicts; by contrast, the South has fewer people in the moderate category and more in the other two categories than the model predicts. This residual analysis suggests that the groups differ in dispersion as well as location, with relatively more dispersion in the South. Section 5.4 presents a cumulative logit model that allows dispersion as well as location effects.

Nonetheless, the estimates from the proportional odds form of model are useful for describing overall location tendencies of the four regions. The estimates convey the fact that fundamentalism is considerably more likely in the South than in other regions and also somewhat more likely in the Midwest than in the Northeast or West/Mountain states.

3.5.9 Model Selection Issues

For a candidate set of potential models, how should we select one? The usual approaches for model selection are available. For example, at the start of a study, we could formulate certain hypotheses to be tested that correspond to comparing

---

[5]Obtained using the mph.fit R function discussed in the Appendix as shown at www.stat.ufl.edu/~aa/ordinal/ord.html.

pairs of nested models. A very small P-value in a test comparing two models suggests rejecting the simpler model in favor of the more complex model, subject to the usual caveat about whether a statistically significant result is also practically significant.

Other criteria besides significance tests can help select a good model. The best known is the *Akaike information criterion* (AIC). It judges a model by how close its fitted values tend to be to the true outcome probabilities, as summarized by a certain expected distance between the two. The estimated optimal model for providing the best estimates is the model that minimizes

$$\text{AIC} = -2(\text{log likelihood} - \text{number of parameters in model}).$$

The AIC penalizes a model for having many parameters. Even though a simple model is farther than a more complex model from the true relationship, for a sample the simple model may provide better estimates of the true expected values.

For Table 3.1 the AIC values are 2664.2 for the quantitative predictor (uniform association) model (3.12) and 2668.4 for the qualitative predictor model (3.14). This criterion favors the simpler model (3.12).

3.6 CUMULATIVE LOGIT MODELS WITHOUT PROPORTIONAL ODDS

A notable feature of the proportional odds form of cumulative logit model is the assumption that the effect of each explanatory variable is the same for the logits for the different cumulative probabilities. An advantage of this model is that effects are simple to summarize and interpret, requiring only a single parameter for each predictor.

In Section 3.3.2 we showed that the proportional odds form of model is implied by a latent variable structure in which an underlying continuous response variable having a logistic conditional distribution satisfies an ordinary regression model with the same dispersion at all predictor values. When this ordinal model fails, the latent variable model usually fails also. Often, this is because either the linear predictor is inadequate (e.g., lacking an important interaction term) or because the dispersion varies substantially among the predictor values.

3.6.1 Cumulative Logit Models with Separate Effects for Each Logit

The generalized model (3.14) that permits different effects of the explanatory variables for the different cumulative logits is

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \beta'_j \mathbf{x}, \qquad j = 1, \ldots, c - 1.$$

As in the baseline-category logit model, each predictor has $c - 1$ parameters. This model implies nonparallelism of the lines for the different cumulative logits and of

the curves for the different cumulative probabilities. Therefore, curves for different cumulative probabilities can cross for some x values. Such curves then violate the proper order for cumulative probabilities. So although this more general model can be useful, it can hold only over a limited range of predictor values. In practice, it is more useful for contingency tables with at most a few variables than for data sets with several predictors, of which some are continuous. (This limitation does not apply to the ordinal logistic models introduced in Chapter 4, which are valid even when different logits have different effects.)

Because cumulative probabilities may be out of order at some settings of the predictors, model fitting can fail for this model. That is, it may not be feasible to maximize the multinomial likelihood function that considers this model simultaneously for all j. When we impose constraints on the cumulative probabilities, model fitting becomes considerably more complex. We can fit the model separately for the different j, but this does not provide an overall maximized likelihood function.

This more general model also has other disadvantages. Although the bias diminishes in estimating the population response proportions at the various predictor settings, the model is much less parsimonious because of the possibly large increase in the number of parameters. The mean-squared errors of the estimates of the population response proportions may even tend to be larger. These comments reflect the trade-off in statistical analysis between bias and variance, both of which contribute to mean-squared error. Simpler models have greater bias, yet may provide better estimates in terms of a criterion such as mean-squared error because of the decreased variance of estimation that results from their parsimony.

3.6.2   Example: Mental Health Modeling Revisited

The example in Section 3.4.4 modeled mental impairment with four response categories as a function of a quantitative life events index and a binary SES indicator $(1 = \text{high}, 0 = \text{low})$ using the cumulative logit model with proportional odds structure,

$$\text{logit } [P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2, \qquad j = 1, 2, 3.$$

The more general model with separate effects for each logit is

$$\text{logit } [P(Y \leq j)] = \alpha_j + \beta_{j1} x_1 + \beta_{j2} x_2, \qquad j = 1, 2, 3.$$

Table 3.10 shows the parameter estimates for this model obtained by fitting an ordinary logistic regression model separately for $j = 1, 2, 3$. The estimated life events and SES effects are substantively similar for the three logits, suggesting that it is simpler to use the proportional odds form of model. The final column of Table 3.10 shows the estimates of parameters for that model. Comparing the standard errors for this model with the other models illustrates the efficiency benefit of using the entire four-category response instead of collapsing it to a binary response.

TABLE 3.10. Estimates for Cumulative Logit Modeling of Mental Impairment Data of Table 3.3 Allowing Different Effects for Each Logit

Parameter	Logit 1 (SE)	Logit 2 (SE)	Logit 3 (SE)	Prop. Odds (SE)
Intercept	−0.173 (0.748)	0.9251 (0.723)	2.595 (0.975)	
Life events	−0.328 (0.164)	−0.3099 (0.148)	−0.376 (0.166)	−0.319 (0.121)
SES	1.006 (0.784)	1.6297 (0.781)	0.947 (0.868)	1.111 (0.611)

3.6.3 Example: Religious Fundamentalism by Region Revisited

In Table 3.8 we cross-classified subjects by region (Northeast, Midwest, South, West/Mountain) and by whether they consider themselves fundamentalist, moderate, or liberal in their religious beliefs. Since region is nominal scale, we created indicator variables for its categories and used the model

$$\text{logit } [P(Y \leq j)] = \alpha_j + \tau_i,$$

with $\tau_4 = 0$. The score test of the proportional odds assumption (Table 3.9) showed that this model fits poorly. For the more general model with separate effects for each logit, fitting the model separately for each logit gives estimates of (τ_1, τ_2, τ_3) equal to $(-0.45, 0.43, 1.15)$ for the first logit and $(0.09, 0.18, 0.58)$ for the second logit. The change in the sign of $\hat{\tau}_1$ reflects the lack of stochastic ordering of the first region (Northeast) and the fourth region (West/Mountain). A Northeast resident is less likely to be fundamentalist, reflected by $\hat{\tau}_1 = -0.45 < 0$ for the first logit, but slightly more likely to be fundamentalist or moderate and hence slightly less likely to be liberal, as reflected by $\hat{\tau}_1 = 0.09 > 0$ for the second logit. Inspection of the conditional distributions in Table 3.8 shows less dispersion in the Northeast than the other regions. The $\{\hat{\tau}_2, \hat{\tau}_3\}$ estimates also differ by a fair amount for the two logits, but the direction of the effect is preserved.

In summary, the proportional odds version of the cumulative logit model does not fit these data well. Its summary estimates of (τ_1, τ_2, τ_3) of $(-0.07, 0.27, 0.89)$ are rather severe summaries of the estimates $(-0.45, 0.43, 1.15)$ for the first logit and $(0.09, 0.18, 0.58)$ for the second logit. In fact, even if the Northeast or West/Mountain region were dropped from the data set, so that the three groups remaining in the sample are stochastically ordered, the test of proportional odds would show lack of fit. However, we've seen that the estimates from that simple model convey the basic information about location: Residents of the South are considerably more likely to be fundamentalist, and there is also somewhat more of a tendency for fundamentalism in the Midwest than in the Northeast or in West/Mountain states. Even though the proportional odds form of model has lack of fit, it is still useful for summarizing overall location effects in the data.

3.6.4 Partial Proportional Odds

Peterson and Harrell (1990) proposed a model that falls between the proportional odds version of the cumulative logit model and the more general model (3.14). In

this model, some predictors in the set $\mathbf{x}$ have a proportional odds structure, but others do not. Denote the subset of predictors that do not have it by $\mathbf{u}$. The *partial proportional odds model* is

$$\text{logit}\,[P(Y \le j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x} + \boldsymbol{\gamma}'_j\mathbf{u}, \qquad j = 1, \ldots, c - 1. \qquad (3.15)$$

For identifiability, one of the $\boldsymbol{\gamma}_j$, say $\boldsymbol{\gamma}_1$, equals $\mathbf{0}$.

   For a predictor x_k having proportional odds, the parameter β_k has the ordinary cumulative log odds ratio interpretation that holds for each of the $c - 1$ cumulative probabilities. For a predictor x_k not having proportional odds, β_k is the log odds ratio only for the first cumulative probability. Denote by u_k the element of $\mathbf{u}$ that is also x_k. Then the conditional log odds ratio between Y and x_k, controlling for the other variables, is $\beta_k + \gamma_{kj}$ for j between 2 and $c - 1$. The ordinary proportional odds model is the special case in which

$$\boldsymbol{\gamma}_2 = \cdots = \boldsymbol{\gamma}_{c-1} = \mathbf{0}.$$

Peterson and Harrell (1990) also proposed special cases of this model in which the parameters for the nonproportional odds part of the model satisfy certain constraints. For example, suppose that predictor x_k is the only one with nonproportional odds, and suppose that the conditional log cumulative odds ratio between Y and x_k changes linearly as the cutpoint changes from 1 to $c - 1$. Let γ_j denote the increment to the effect β_k of x_k for cumulative probability j. Then we could consider the special case of this model in which $\gamma_j = (j - 1)\gamma$. The advantage of such a model is that it has only one more parameter than the proportional odds version of the model.

3.6.5   Example: Coronary Heart Disease and Smoking

Table 3.11 from Peterson and Harrell (1990) shows the relationship between the degree of coronary heart disease and smoking status in a study at Duke University Medical Center. Let x be an indicator for smoking status ($x = 1$ smoker, $x = 0$ nonsmoker). The proportional odds form of cumulative logit model,

$$\text{logit}\,[P(Y \le j)] = \alpha_j + \beta x,$$

has $\hat{\beta} = -0.737$ (SE $= 0.082$). The estimated common cumulative odds ratio is $e^{-0.737} = 0.48$. However, the model fits poorly, with the goodness-of-fit tests having $X^2 = 40.3$ and $G^2 = 40.5$ and the score test of the proportional odds assumption having test statistic 44.8 (df $= 3$). These tests all have an alternative hypothesis corresponding to the general model with three additional parameters (i.e., β replaced by β_j), which is saturated.

   The lack of fit is reflected by the four sample cumulative log odds ratios, which equal -1.04, -0.65, -0.46, and -0.07. In other words, a strong association occurs when the outcome is measured as "no disease" versus "some disease," but the association weakens progressively to nearly no association when the outcome is

TABLE 3.11. Smoking Status and Degree of Heart Disease, with Percentages for Degree of Heart Disease in Parentheses

Smoking	Degree of Coronary Heart Disease[a]				
Status	0	1	2	3	4
Smoker	350 (22.6%)	307 (19.8%)	345 (22.3%)	481 (31.0%)	67 (4.3%)
Nonsmoker	334 (45.2%)	99 (13.4%)	117 (15.8%)	159 (21.5%)	30 (4.1%)

Source: Peterson and Harrell (1990), with permission.
[a]0, no disease; 4, very severe disease.

measured by contrasting the most severe disease category with the others. There seems to be roughly a decreasing linear trend in these cumulative log odds ratios.

We consider next the model

$$\text{logit} \, [P(Y \leq j)] = \alpha_j + \beta_1 x + (j - 1)\beta_2 x.$$

This is not a proportional odds model, because the effect of x depends on j. The cumulative log odds ratio contrasting outcomes $Y \leq j$ and $Y > j$ is

$$\log \theta_{11}^C = \beta_1, \quad \log \theta_{12}^C = \beta_1 + \beta_2, \quad \log \theta_{13}^C = \beta_1 + 2\beta_2, \quad \log \theta_{14}^C = \beta_1 + 3\beta_2.$$

This model has a much better fit ($X^2 = 3.45$, $G^2 = 3.43$, df $= 2$, P-value $= 0.18$). The ML estimates of the effect parameters[6] are $\hat{\beta}_1 = -1.017$ (SE $= 0.094$) and $\hat{\beta}_2 = 0.298$ (SE $= 0.047$). The corresponding estimated cumulative log odds ratios are

$$\log \hat{\theta}_{11}^C = -1.02, \quad \log \hat{\theta}_{12}^C = -0.72, \quad \log \hat{\theta}_{13}^C = -0.42, \quad \log \hat{\theta}_{14}^C = -0.12.$$

These represent well the sample values of -1.04, -0.65, -0.46, and -0.07.

3.6.6 Other Approaches When Proportional Odds Fits Poorly

When a proportional odds model does not fit adequately, what are the remedies? We have seen two possibilities: using a more general model such as the nonproportional odds model (3.14) or the partial proportional odds model (3.15). Sometimes we can continue to use the proportional odds model for describing the essence of the location effects, as shown for the example in Section 3.6.3. We list some other options here briefly and discuss them further in subsequent chapters.

Most options involve other types of models having additional parameters. These options include (1) using a link function for which the response curve is nonsymmetric (Section 5.3); (2) adding additional terms, such as interactions, to the linear predictor; (3) adding dispersion parameters (Section 5.4); (4) fitting logistic models

---

[6]Obtained using the mph.fit R function described in the Appendix as shown at www.stat.ufl.edu/~aa/ordinal/ord.html.

that have separate parameters for each logit; and (5) letting the cutpoint parameters $\{\alpha_j\}$ depend on covariates through a linear model (Terza 1985).

Option 2 is usually worth investigation because the variability across the $c - 1$ cumulative logits for the effect of a particular predictor may reflect an interaction between that predictor and another one. Failure of the proportional odds model often reflects nonconstant variability of the response variable, but in practice, option 3 has seen relatively little use. For option 4, other than fitting an ordinary logistic regression model separately to each cumulative logit, one could use a different type of ordinal logit. For example, Section 4.2 presents the continuation-ratio logit model. This model utilizes the ordinal nature of Y but can have different effects for each of the $c - 1$ logits. Section 4.1 presents the adjacent-categories logit model, which in its most general form is equivalent to a standard baseline-category logit model (Section 4.1.3). This general form treats the response variable as nominal, but one can use the ordinality in an informal way in interpreting the effects and how they may tend to increase or decrease across the $c - 1$ adjacent-categories logits. An advantage of these types of logits is that the ML fit of a model permitting heterogeneous effects, found simultaneously for all $c - 1$ logits, exists much more generally that it does for the cumulative logit model (3.14) with heterogeneous effects.

3.7 CONNECTIONS WITH NONPARAMETRIC RANK METHODS

We've seen that the effect of an explanatory variable on an ordinal response variable can be tested by constructing a cumulative logit model and testing the hypothesis that a certain parameter or set of parameters equals zero. When the model has a proportional odds structure, such tests have connections with nonparametric tests using ranks or rank-based scores for the response variable.

3.7.1 Connections with Comparing Groups Using Mean Ranks

Consider the comparison of two groups on an ordinal response Y. The Wilcoxon test is a nonparametric method for comparing two groups by ranking all the observations on the response variable and comparing the mean ranks. In Section 7.4.1 we show that the Wilcoxon test generalizes to allow tied observations, as occur with ordered categorical data. The data are summarized in a $2 \times c$ contingency table of counts $\{n_{ij}\}$ with ordered columns, such as for Table 3.11 on smoking status and $Y =$ degree of coronary heart disease. The test compares means using midrank scores $\{r_j\}$ or corresponding ridit scores $\{a_j\}$ based on the marginal proportions $\{p_j = (n_{1j} + n_{2j})/n\}$ of Y, as defined in Section 2.1.1. The null hypothesis of identical response probabilities for the two groups is tested by comparing the mean ranks for the two groups, relative to the variability expected under the null hypothesis.

The Wilcoxon test has a connection with a cumulative logit model–based test. For $2 \times c$ tables with ordered columns, we regard the data as two independent multinomial observations, with sample sizes n_1 and n_2 for the two rows. The model with proportional odds structure is

$$\text{logit } [P(Y \leq j)] = \alpha_j + \beta x_i, \tag{3.16}$$

where x_i is an indicator variable for the groups (rows) that equals 1 in row 1 ($i = 1$) and 0 in row 2 ($i = 2$). We can test the hypothesis of identical response distributions for the two rows by a Wald, likelihood-ratio, or score test of H_0: $\beta = 0$. The score test is based on the derivative of the log likelihood with respect to β evaluated at the ML estimates under the constraint $\beta = 0$. This derivative is proportional to

$$\sum_{i=1}^{2} \sum_{j=1}^{c} (n_{ij} - n_i p_j) x_i r_j$$

(McCullagh and Nelder 1989, p. 188). This is the difference between the sum of rank scores in group 1 and its null expected value for the two groups. Equivalently, it can be expressed in terms of the difference between the mean ranks for the two groups. In fact, the score test is equivalent to the discrete version of the Wilcoxon test. Such a test is locally most powerful for the one-sided alternative hypothesis (McCullagh 1980, Sec. 4.3). Pettitt (1984a) considered connections for underlying distributions other than the logistic.

The more general qualitative predictor model (3.14) for comparing r groups in a $r \times c$ table is

$$\text{logit } [P(Y \leq j)] = \alpha_j + \tau_i, \qquad j = 1, \ldots, c - 1.$$

The score test of H_0: $\tau_1 = \tau_2 = \cdots = \tau_r$ compares the mean ranks of the r rows, with df $= r - 1$. It is equivalent to the generalized Kruskal–Wallis test for an ordered categorical response presented in Section 7.4.3.

Finally, for an arbitrary cumulative logit model of proportional odds form, consider a particular explanatory variable k with value x_{ik} for setting i of the explanatory variables and parameter β_k. Then the derivative of the log-likelihood function with respect to β_k uses the data in terms of $\sum_i (\sum_j x_{ik} n_{ij} r_j)$, where n_{ij} is the number of observations in response category j at setting i of the explanatory variables. For example, if explanatory variable k is binary and its indicator values are $+1$ and -1, inference about the effect of that predictor essentially uses the difference between the mean ranks for the two levels of the predictor.

3.7.2 Sample Size and Power for Comparing Two Groups

Whitehead (1993) presented sample-size formulas for achieving a desired power in comparing two groups on an ordered categorical response using the proportional odds version (3.16) of the cumulative logit model with a binary indicator predictor. Let $1 - \beta$ denote the power for an α-level test for detecting an effect of size β_0 for the cumulative log odds ratio in that model. Suppose that the plan is to allocate the sample size to the two groups in the ratio A to 1, and π_j denotes the guess for the marginal probability in response category j. Based on large-sample approximations, the sample size required for a two-sided test is approximately

$$n = \frac{3(A + 1)^2 (z_{\alpha/2} + z_\beta)^2}{A \beta_0^2 (1 - \sum \pi_j^3)}.$$

This requires anticipating the marginal probabilities and the size of the effect. Alternative approximations are mentioned in Note 3.9.

Setting $\{\pi_j = 1/c\}$ provides a lower bound for n. Whitehead showed that the sample size does not depart much from this bound unless a single dominant response category occurs. With $\{\pi_j = 1/c\}$, the needed n depends on c through the proportionality constant $(1 - 1/c^2)^{-1}$. In this sense, relative to a continuous response ($c = \infty$), using c categories provides efficiency $(1 - 1/c^2)$; for $c = (2, 3, 4, 5, 10)$, this is $(0.75, 0.89, 0.94, 0.96, 0.99)$. The loss of information from collapsing to a binary response is substantial, but there is little gain from using more than four categories. The ratio of the sample size $n(c)$ needed for c categories relative to the sample size $n(2)$ needed when $c = 2$ is

$$\frac{n(c)}{n(2)} = 0.75 \left(1 - \frac{1}{c^2}\right)^{-1}.$$

However, this assumes equal dispersion among the categories, which is not usual in practice.

3.7.3 Testing Conditional Independence in Three-Way Tables

As we discussed in Section 2.4, in practice we usually study the effect of an explanatory variable on a response variable while controlling for other variables. Models provide a natural way to do this. For an ordinal response variable, we can construct a model for which conditional independence between an explanatory variable and the response corresponds to a value of 0 for a model parameter. The test of the hypothesis then uses standard methods for testing that a parameter (or set of parameters) equals 0.

For cumulative logit modeling of Y, we consider two cases, differing in terms of whether the explanatory variable X is treated as quantitative (i.e., interval scale, or ordinal scale with monotone scores) or as qualitative (nominal scale). In each case, the control variable Z, which could be multivariate, is treated as nominal. A partial table relates X and Y at each level of Z. When the XY association is similar in the partial tables, the power benefits from basing a test statistic on a model of homogeneous conditional association. Alternatively, in Section 6.4.5 we present score tests that generalize the Cochran–Mantel–Haenszel (CMH) test for sets of 2×2 tables to sets of $r \times c$ tables.

X Quantitative   Let $\{u_i\}$ be ordered scores for the rows of a contingency table. The model

$$\text{logit } [P(Y \leq j \mid X = i, Z = k)] = \alpha_j + \beta u_i + \beta_k^Z \tag{3.17}$$

and the more general model with $\alpha_j + \beta_k^Z$ replaced by α_{jk} have the same linear trend β for the effect of X in each partial table. The models also apply when X is continuous, in which case u_i is the value of X at its ith level. For these models, XY

conditional independence is H_0: $\beta = 0$. Likelihood-ratio, score, or Wald statistics for H_0 provide large-sample chi-squared tests with df $= 1$ that are sensitive to the trend alternative. The score test cumulates correlation-type information across the partial tables, with scores $\{u_i\}$ for the rows and midrank scores (or, equivalently, ridit scores) for Y.

X Qualitative   Alternatives to conditional independence that treat X as a nominal-scale factor are

$$\text{logit } [P(Y \le j \mid X = i, Z = k)] = \alpha_j + \beta_i^X + \beta_k^Z,$$

and the more general model with $\alpha_j + \beta_k^Z$ replaced by α_{jk}. The effect parameters have a constraint such as $\beta_r^X = 0$. For these models, XY conditional independence is H_0: $\beta_1^X = \cdots = \beta_r^X$. Large-sample chi-squared tests have df $= r - 1$. The score test cumulates information about the common variability among the row mean ranks on Y across the partial tables, with average rank scores for Y. It is a stratified version of the Kruskal–Wallis test (Section 7.4.3).

3.7.4   Example: Political Ideology and Evolution, by Religiosity

The 2006 General Social Survey shows a moderate association between political ideology (measured on a scale of 1 to 7, from extremely liberal to extremely conservative) and whether one believes that human beings evolved from earlier species of animals. Using an indicator variable for belief in evolution (1 = yes, 0 = no), model (3.16) with political ideology as the response variable has $\hat{\beta} = 0.908$ (SE $= 0.094$). The estimated common cumulative odds ratio is $\exp(0.908) = 2.48$, those who believe in evolution being relatively more liberal. Could this association be explained by religiosity, with more religious subjects tending to be more conservative and also less likely to believe in evolution? We'll use this GSS to test conditional independence.

    Table 3.12 cross-classifies these three variables. Using two indicator variables for the religiosity effect, model (3.17) has $\hat{\beta} = 0.697$ (SE $= 0.100$). The association seems somewhat weaker than when religiosity is ignored, but the likelihood-ratio statistic for testing H_0: $\beta = 0$ equals 49.7 (df $= 1$). There is very strong evidence that belief in evolution tends to be associated with more liberal political beliefs, even after controlling for religiosity.

    For the religiosity effect, the likelihood-ratio statistic equals 43.2 (df $= 2$). For coding that sets $\hat{\beta}_3^Z = 0$, model (3.17) has $\hat{\beta}_1^Z = 0.754$ and $\hat{\beta}_2^Z = 0.519$. Lower religious attendance tends to be associated with a tendency toward more liberal views, for given beliefs about evolution. Using, instead, a linear trend for religiosity, for scores (1, 2, 3) for religiosity, the estimated coefficent for the religiosity effect is -0.369 (SE $= 0.058$), and the likelihood-ratio statistic for that effect is 41.4 (df $= 1$).

    These models assume a lack of interaction between religiosity and belief in evolution in their effects on political ideology. Adding an interaction term to the

TABLE 3.12. Data on Political Ideology and Belief in Evolution, by How Often Attend Religious Services

Religiosity[b]	Evolution	Political Ideology[a]						
		1	2	3	4	5	6	7
1	Yes	23	83	66	161	67	33	9
	No	8	22	16	108	24	34	4
2	Yes	8	34	30	68	36	24	1
	No	4	15	19	64	30	38	5
3	Yes	5	15	15	48	19	18	3
	No	4	17	36	113	51	113	37

[a] 1, extremely liberal; 2, liberal; 3, slightly liberal; 4, moderate; 5, slightly conservative; 6, conservative; 7, extremely conservative.
[b] 1, at most once a year; 2, several times a year to two or three times a month; 3, nearly every week or more.

model having two indicator variables for religiosity gives two extra parameters. The evidence of interaction is not strong (likelihood-ratio statistic = 4.51, df = 2, P-value = 0.10).

Incidentally, these models all show some lack of fit. For example, model (3.17) with no interaction term has $X^2 = 56.8$ and $G^2 = 57.2$ for testing goodness of fit (df = 27, P-value = 0.001). Nonetheless, they are adequate for providing strong evidence of effects. As Mantel (1963) argued in a similar context: "that a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that a test of a linear component of regression provides power for detecting any progressive association which may exist." To illustrate, if we fit the more general version of model (3.17) with $\alpha_j + \beta_k^Z$ replaced by α_{jk}, the model fits somewhat better, with $X^2 = 31.2$ and $G^2 = 32.0$ for testing goodness of fit (df = 17). However, the amount of evidence about the evolution effect is very similar, as $\hat{\beta} = 0.692$ (SE = 0.100) and the likelihood-ratio statistic for testing H_0: $\beta = 0$ equals 48.8 (df = 1). In this case, the statistics detect the tendency of those who believe in evolution to be more liberal and for those who are more religious to be less liberal.

CHAPTER NOTES

Section 3.1: Types of Logits for an Ordinal Response

3.1. McCullagh (1978) defined a statistical method to be *palindromic invariant* if the results are invariant under a complete reversal of order of categories (except for the sign of the estimates) but not under general permutations of categories. Most methods discussed in this book, such as cumulative logit modeling, have this property for ordinal variables. A reversal of the categories of Y changes the sign of parameter estimates but does not alter the maximized log-likelihood or substantive

conclusions. Models presented in Section 4.2 using continuation-ratio logits are not palindromic invariant.

3.2. This book focuses on methods for ordinal response variables. Other methods are designed specially for ordinal explanatory variables, for various types of response variables (see Sections 7.4.8 and 7.5.1). The first of these sections presents the *Cochran–Armitage trend test* for a binary response. It assigns scores to the categories of an ordinal explanatory variable as the basis of a chi-squared statistic with df $= 1$ for detecting evidence of a trend in the probability of a particular outcome (Armitage 1955; Cochran 1954, 1955).

Section 3.2: Cumulative Logit Models

3.3. Cumulative logit models were proposed for contingency tables by many authors, including Snell (1964), Bock and Jones (1968), Samejima (1969), Tukey (1971), Williams and Grizzle (1972), Simon (1974), Clayton (1974), and Bock (1975, pp. 544–546). McCullagh (1980) popularized the proportional odds case. His influential article and the subsequent discussion gave an interesting exposition of issues related to modeling ordinal data. McCullagh also presented more general models with arbitrary link functions and/or dispersion terms (Sections 5.1 and 5.4). Later articles about cumulative logit models include McCullagh (1984), Snapinn and Small (1986), Stram et al. (1988), Hastie et al. (1989), Tutz (1989), Brant (1990), Peterson and Harrell (1990), Holtbrügge and Schumacher (1991), Agresti and Lang (1993a), Joffe and Greenland (1995), Scott et al. (1997), and Marshall (1999). See also Sections 8.2.1, 8.2.4, 8.6.1, 8.4.2, 10.1.1, 10.3, and notes in Chapters 8, 9, and 10 for their use in models for clustered observations, such as in longitudinal studies, and Section 11.3 for a Bayesian approach. In the context of survival modeling, see Clayton (1976), Bennett (1983), Pettitt (1984b), Rossini and Tsiatis (1996), and Hedeker et al. (2000). For smoothing and semiparametric structuring using cumulative logit models, see Hastie and Tibshirani (1987), Yee and Wild (1996), Fahrmeir and Tutz (2001), Kauermann and Tutz (2003), and Tutz (2003). For example, in Kauermann and Tutz (2003), some explanatory variables enter the model linearly, whereas others have unspecified but smooth functions.

Section 3.4: Fitting and Inference for Cumulative Logit Models

3.4. Kaufmann (1988) discussed existence and uniqueness of ML estimates for ordinal response models, including cumulative logit models and continuation-ratio logit models and models with other link functions. When all cell counts are positive, he showed that the ML estimates exist and they are unique if all parameters are identifiable.

3.5. With highly stratified data, the number of parameters can be large relative to the sample size, and unconditional ML estimates may perform poorly. A standard approach with canonical-link GLMs such as binary logistic regression conditions on sufficient statistics for nuisance parameters, thus eliminating those parameters from the conditional likelihood function. This approach does not work for cumulative logit models, because baseline-category logits rather than cumulative logits are

the canonical link for a multinomial distribution. McCullagh (1984) proposed a type of sequential conditioning as a way of eliminating nuisance parameters from cumulative logit models.

3.6. For other articles about ordinal classification, see Larichev and Moshkovich (1994), Rudolfer et al. (1995), Coste et al. (1997), Feldman and Steudel (2000), Tutz and Hechenbichler (2005), Horvath and Vojtas (2006), and Piccarreta (2008). Tutz and Hechenbichler (2005) used variants of bagging and boosting methods that make use of the ordinality, showing how predictive power improves with appropriate aggregation. In the context of machine learning, see also Herbrich et al. (1999), Frank and Hall (2001), Shashua and Levin (2003), Chu and Ghahramani (2005), Chu and Keerthi (2007), and Waegemana et al. (2008).

Section 3.6: Cumulative Logit Models Without Proportional Odds

3.7. Cumulative logit models without the proportional odds structure were proposed by Williams and Grizzle (1972), who used weighted least squares for model fitting. Other articles that considered models with different effect parameters for different logits include Cox and Chuang (1984), Brant (1990), Peterson and Harrell (1990), Cox (1995), Ananth and Kleinbaum (1997), Bender and Groven (1998), Lall et al. (2002), and Cole et al. (2004). For alternative tests of the proportional odds property, see Brant (1990) and Stiger et al. (1999). For further discussion of the partial proportional odds form of model, see Peterson and Harrell (1990), Stokes et al. (2000, Sec. 15.13), Lall et al. (2002), and criticism by Cox (1995).

3.8. Tutz (1989) proposed a compound model of hierarchical form defined on a partition of the response categories into sets. The first part of the model describes the mechanism for classification in those sets, and the second part of the model describes classification into categories within sets. This is useful when groups of categories are relatively homogeneous. For instance, with $c = 7$ categories, a cumulative logit model might describe effects of explanatory variables on whether the response is in category set 1–2, 3–5, or 6–7, and then logistic and cumulative logit models with different parameter values might describe the effects of explanatory variables conditional on the response being in a particular set of categories.

Section 3.7: Connections with Nonparametric Rank Methods

3.9. For determining sample size for comparing two groups, Kolassa (1995) provided methods based on a Cornish–Fisher approximation to the null distribution and an Edgeworth approximation for the power. Hilton and Mehta (1993) provided another approach, based on evaluating the exact conditional distribution with a network algorithm, or simulating that distribution. Rabbee et al. (2003) presented simpler approximations for exact methods. Lee et al. (2002) evaluated methods and claimed that Whitehead's formula is adequate when the sample size is moderately large.

EXERCISES

3.1. Is a cumulative logit model a special case of a baseline-category logit model? Why not?

3.2. For the model, logit $[P(Y \leq j)] = \alpha_j + \beta_j x$, explain why cumulative probabilities may have an inappropriate order for some x values.

3.3. A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, don't know). One way to model such a scale uses an ordinary binary logistic model for the probability of a don't know response and uses a separate ordinal model for the four ordered categories conditional on response in one of those categories. Explain how to construct a likelihood function to do this simultaneously. See also Note 3.8.

3.4. Consider the $2 \times 3 \times 2$ contingency table relating a binary x_1 to an ordinal Y, by row, of (10, 10, 10; 0, 0, 30) at the first level of x_2 and (10, 10, 10; 10, 10, 10) at the second level of x_2. Explain why a cumulative logit model with main effects has finite estimates for the effects of x_1 and x_2 but a model that also has an interaction term does not. Observe what happens when you fit these models with software.

3.5. Table 3.13 shows data cross-classifying job satisfaction and income, stratified by gender, for black Americans. The data are analyzed in Agresti (2002, Chap. 7).

    (a) Analyze the conditional association between job satisfaction and income, treating the variables as nominal.

    (b) Analyze the conditional association between job satisfaction and income, treating the variables as ordinal.

    (c) Compare the analyses in parts (a) and (b) in terms of simplicity of description and power for detecting an association.

TABLE 3.13. Job Satisfaction and Income by Gender

Income[b]	Females' Job Satisfaction[a]				Males' Job Satisfaction[a]			
	1	2	3	4	1	2	3	4
1	1	3	11	2	1	1	2	1
2	2	3	17	3	0	3	5	1
3	0	1	8	5	0	0	7	3
4	0	2	4	2	0	1	9	6

Source: General Social Survey.
[a] 1, very dissatisfied; 2, a little satisfied; 3, moderately satisfied; 4, very satisfied.
[b] 1, lowest; 4, highest.

3.6. Refer to Exercise 2.7. Analyze these data using methods of this chapter.

CHAPTER 4

Other Ordinal Logistic Regression Models

In Chapter 3 we modeled ordinal responses using logits for cumulative probabilities. In this chapter we present alternative logit models using the *adjacent-categories logits* and the *continuation-ratio logits* introduced in Section 3.1. Such models have interpretations that can use individual categories rather than the cumulative probabilities. Like the proportional odds version of the cumulative logit model, the proportional odds versions of the adjacent-categories logit model and the continuation-ratio logit model account for ordinality by assuming that explanatory variables have the same effects for each of the $c - 1$ logits for a c-category response variable. More generally, models with these types of logits, unlike cumulative logit models, have a valid structure (e.g., cumulative probabilities maintaining the appropriate order) even when used with separate effect parameters for each logit.

In Section 4.3 we present a more general type of proportional odds model for adjacent-category logits or corresponding baseline-category logits. Called the *stereotype model*, it estimates scores for the response categories as a way of permitting more general structure for the effects while maintaining a similar effect structure for each logit. It differs from other models considered so far in having a predictor form that is multiplicative rather than linear in the parameters.

4.1 ADJACENT-CATEGORIES LOGIT MODELS

For multinomial probabilities $\{\pi_j\}$, the adjacent-categories logits are

$$\text{logit}\,[P(Y = j \mid Y = j \text{ or } Y = j + 1)] = \log \frac{\pi_j}{\pi_{j+1}}, \qquad j = 1, \ldots, c - 1.$$

With a set of explanatory variables **x**, the general adjacent-categories logit model has the form

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}, \qquad j = 1, \dots, c - 1. \tag{4.1}$$

Because such models use pairs of adjacent categories, the effects are naturally described with local odds ratios rather than the cumulative odds ratios that naturally apply with cumulative logit models. Unlike the corresponding cumulative logit model (3.14) with nonproportional odds presented in Sections 3.5.5 and 3.6.1, this model provides valid probabilities regardless of predictor values.

4.1.1 Proportional Odds for Adjacent-Categories Logit Models

The construction of the adjacent-categories logits recognizes the ordering of the categories of Y. To benefit from this in model parsimony by truly exploiting the ordinality of Y, however, we must use a simpler specification for the linear predictor. If an explanatory variable has a similar effect for each logit, the usual advantages of model parsimony accrue from using a single parameter instead of $c - 1$ parameters to describe that effect. The model is then

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}' \mathbf{x}, \qquad j = 1, \dots, c - 1. \tag{4.2}$$

For predictor k, the estimated odds of the lower instead of the higher of two adjacent response categories multiply by $\exp(\beta_k)$ for each 1-unit increase in x_k. This odds ratio is the same for all adjacent pairs, that is, not dependent on j. The corresponding model for the category probabilities is

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x})}{1 + \sum_{k=1}^{c-1} \exp(\alpha_k + \boldsymbol{\beta}' \mathbf{x})}, \qquad j = 1, 2, \dots, c - 1.$$

Model (4.2) has *proportional odds* structure, much like the corresponding cumulative logit model (3.6). These two types of model fit well in similar situations. One reason for this is that they both imply stochastically ordered distributions for Y at different predictor values. A model more general than this but simpler than model (4.1) permits *partial proportional odds*, having simpler structure for some but not all explanatory variables. Such a model is an analog of the partial proportional odds model (3.15) for cumulative logits.

For the parameterization (4.2), a value for $\beta_k > 0$ means that as x_k increases, Y is more likely to fall at lower values. For the sign of β_k to have the usual interpretation by which a positive value means a positive conditional effect of x_k on Y, the model can instead be expressed as

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j - \boldsymbol{\beta}' \mathbf{x}.$$

This parallels the parameterization logit $P(Y \leq j) = \alpha_j - \boldsymbol{\beta}'\mathbf{x}$ often used for the proportional odds version of the cumulative logit model [see equation (3.8)]. Another way to have such an interpretation for β_k is to express the model in terms of $\log[\pi_{j+1}(\mathbf{x})/\pi_j(\mathbf{x})]$, with the probability for the higher category in the numerator instead of the denominator.

4.1.2 Parallel Log Odds Models for $r \times c$ Tables

Consider an $r \times c$ contingency table with an ordinal response variable Y. Denote the conditional probabilities $\{\pi_{j|i} = P(Y = j \mid X = i)\}$.

First, suppose that X is a quantitative or ordinal explanatory variable. Let $\{u_i\}$ denote ordered row scores for X. A model of proportional odds form that utilizes the ordering of the rows is

$$\log \frac{\pi_{j|i}}{\pi_{j+1|i}} = \alpha_j + \beta u_i, \qquad j = 1, \ldots, c - 1. \tag{4.3}$$

For this model, the local log odds ratio satisfies

$$\log \theta_{ij}^L = \beta(u_i - u_{i+1}).$$

For equally spaced scores, the model satisfies *uniform association* for the local odds ratio. When all $u_i - u_{i+1} = 1$, the uniform local odds ratio equals $\exp(\beta)$. Cumulative logit model (3.9) is a corresponding uniform association model for cumulative odds ratios.

For a nominal explanatory variable, we replace the ordered $\{\beta u_i\}$ by unordered parameters $\{\tau_i\}$. This results in the more general *row effects* model, proposed by Simon (1974),

$$\log \frac{\pi_{j|i}}{\pi_{j+1|i}} = \alpha_j + \tau_i, \qquad j = 1, \ldots, c - 1, \tag{4.4}$$

with a constraint such as $\tau_r = 0$. The local log odds ratio satisfies

$$\log \theta_{ij}^L = \tau_i - \tau_{i+1}.$$

For a given pair of rows i and k, the $c - 1$ log odds ratios that are local in the response variable are identical,

$$\log \frac{\pi_{j|i}/\pi_{j+1|i}}{\pi_{j|k}/\pi_{j+1|k}} = \tau_i - \tau_k, \qquad j = 1, 2, \ldots, c - 1.$$

This model is also useful when the explanatory variable is ordinal but we do not expect an overall positive trend or negative trend for the association.

For the quantitative predictor model (4.3) with fixed j, the r logits $\{\log(\pi_{j|i}/\pi_{j+1|i}), i = 1, \ldots, r\}$ plotted against $\{u_i, i = 1, \ldots, r\}$ follow a straight line with slope β. Forming such a plot for each of the $c - 1$ possible values of j yields $c - 1$ parallel lines. By contrast, for the row effects model (4.4), $\{\log(\pi_{j|i}/\pi_{j+1|i}), i = 1, \ldots, r\}$ plotted against $\{i = 1, \ldots, r\}$ are parallel for different j but do not

follow a straight line. Goodman (1983) referred to this model as a *parallel log-odds model*.

4.1.3 Connection with Baseline-Category Logit Models

For nominal-scale response variables, the standard logits are the baseline-category logits. The order of the response categories is then irrelevant, and we contrast an arbitrary baseline category against each of the other categories. When category c is the baseline category, these logits are

$$\log \frac{\pi_1}{\pi_c}, \log \frac{\pi_2}{\pi_c}, \ldots, \log \frac{\pi_{c-1}}{\pi_c}.$$

As noted in Section 3.1.2, the adjacent-categories logits are a basic set of logits that are equivalent to the baseline-category logits. For baseline category c,

$$\log \frac{\pi_j}{\pi_c} = \log \frac{\pi_j}{\pi_{j+1}} + \log \frac{\pi_{j+1}}{\pi_{j+2}} + \cdots + \log \frac{\pi_{c-1}}{\pi_c}. \tag{4.5}$$

Models using adjacent-categories logits can be expressed as baseline-category logit models. For the general adjacent-categories logit model,

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}, \qquad j = 1, \ldots, c - 1,$$

from adding $c - j$ terms as in (4.5), the equivalent baseline-category logit model is

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \sum_{k=j}^{c-1} \alpha_k + \left(\sum_{k=j}^{c-1} \boldsymbol{\beta}'_j \right) \mathbf{x}, \qquad j = 1, \ldots, c - 1$$

$$= \alpha_j^* + \boldsymbol{\beta}_j^{*\prime} \mathbf{x}, \qquad j = 1, \ldots, c - 1.$$

This model has the form of an ordinary baseline category logit model. Because it does not assume a common effect for each j, this model does not utilize the ordinality of Y.

Of greater interest for ordinal responses is the proportional odds form of the adjacent-categories logit model, with common effect $\boldsymbol{\beta}$ for each logit,

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}' \mathbf{x}, \qquad j = 1, \ldots, c - 1.$$

The equivalent baseline-category logit model is

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \sum_{k=j}^{c-1} \alpha_k + (c - j)\boldsymbol{\beta}' \mathbf{x}, \qquad j = 1, \ldots, c - 1 \tag{4.6}$$

$$= \alpha_j^* + \boldsymbol{\beta}' \mathbf{u}_j, \qquad j = 1, \ldots, c - 1, \tag{4.7}$$

with $\mathbf{u}_j = (c - j)\mathbf{x}$. So the adjacent-categories logit model corresponds to a baseline-category logit model with an adjusted model matrix. That model accounts for the ordinality of Y by using a single parameter for each explanatory variable and by letting the explanatory variable itself incorporate a distance measure $c - j$ between each category j and the baseline category c.

The connection between adjacent-categories logits and baseline-category logits is useful because software is more readily available for fitting baseline-category logit models. For example, since

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} - \log \frac{\pi_{j+1}(\mathbf{x})}{\pi_c(\mathbf{x})}, \qquad j = 1, \ldots, c - 1,$$

we can obtain the ML estimate $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}_j$ in the general adjacent-categories logit model (4.1) by obtaining estimates $\{\hat{\boldsymbol{\beta}}_j^*\}$ for the ordinary baseline-category logit model and then finding that

$$\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_j^* - \hat{\boldsymbol{\beta}}_{j+1}^*,$$

where $\hat{\boldsymbol{\beta}}_c^* = 0$.

Similarly, with some software it is possible to fit the proportional odds form of model (4.2). This requires being able to fit the equivalent baseline-category logit model (4.7) that constrains $\{\boldsymbol{\beta}_j\}$ to be identical.[1]

4.1.4   Likelihood Function for Adjacent-Categories Logit Model

The baseline-category logits are the canonical link functions for a multinomial distribution. Unlike models with other link functions, such as cumulative logits and probits, models using them or adjacent-categories logits have reduced sufficient statistics and relatively simple likelihood equations.

For subject i, let y_{ij} denote the binary indicator for whether the response is in category j ($1 = $ yes, $0 = $ no), and let x_{ik} denote the value of explanatory variable k, with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots)'$. Assuming n independent multinomial observations, the contribution of subject i to the log-likelihood function is

$$\log \left[\prod_{j=1}^{c} \pi_j(\mathbf{x}_i)^{y_{ij}} \right] = \sum_{j=1}^{c-1} y_{ij} \log \pi_j(\mathbf{x}_i) + \left(1 - \sum_{j=1}^{c-1} y_{ij} \right) \log \pi_c(\mathbf{x}_i)$$

$$= \sum_{j=1}^{c-1} y_{ij} \log \frac{\pi_j(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)} + \log \pi_c(\mathbf{x}_i).$$

For the baseline-category logit model with parameters α_j^* and $\boldsymbol{\beta}_j^*$ for logit j, the log-likelihood function incorporating all n observations is

---

[1]For example, using PROC CATMOD in SAS or the mph.fit R function, as the Appendix shows.

$$\log \prod_{i=1}^{n}\left[\prod_{j=1}^{c}\pi_j(\mathbf{x}_i)^{y_{ij}}\right]$$

$$= \sum_{i=1}^{n}\left\{\sum_{j=1}^{c-1}y_{ij}(\alpha_j^* + \boldsymbol{\beta}_j^{*'}\mathbf{x}_i) - \log\left[1 + \sum_{j=1}^{c-1}\exp(\alpha_j^* + \boldsymbol{\beta}_j^{*'}\mathbf{x}_i)\right]\right\}$$

$$= \sum_{j=1}^{c-1}\left[\alpha_j^*\left(\sum_{i=1}^{n}y_{ij}\right) + \sum_{k}\beta_{jk}^*\left(\sum_{i=1}^{n}x_{ik}y_{ij}\right)\right]$$

$$- \sum_{i=1}^{n}\log\left[1 + \sum_{j=1}^{c-1}\exp(\alpha_j^* + \boldsymbol{\beta}_j^{*'}\mathbf{x}_i)\right].$$

Now, for the adjacent-categories logit model (4.2) of proportional odds form, because of the connection (4.7) with baseline-category logit models, the log-likelihood function simplifies to

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{c-1}\left[\sum_{k=j}^{c-1}\alpha_k\left(\sum_{i=1}^{n}y_{ij}\right) + \sum_{k}(c-j)\beta_k\left(\sum_{i=1}^{n}x_{ik}y_{ij}\right)\right]$$

$$- \sum_{i=1}^{n}\log\left\{1 + \sum_{j=1}^{c-1}\exp\left[\sum_{k=j}^{c-1}\alpha_k + (c-j)\left(\sum_{k}\beta_k x_{ik}\right)\right]\right\}.$$

The sufficient statistic for α_j is $\sum_{i=1}^{n}\sum_{k=1}^{j}y_{ik}$. This equals the jth cumulative marginal total for Y. The sufficient statistic for β_k is $\sum_i\sum_j x_{ik}(c-j)y_{ij}$. For example, suppose that there is a single explanatory variable x that is ordinal and we apply model (4.3) having the same linear trend for each pair of adjacent categories to a contingency table of counts $\{n_{hj}\}$. Then for the equally spaced row scores $\{u_h = h\}$ for x, the sufficient statistic for β reduces to $\sum_h\sum_j h(c-j)n_{hj}$. Since the row marginal totals on x are fixed by the multinomial sampling design, it is equivalent to know $\sum_h\sum_j(hj)n_{hj}$. For fixed row and column marginal totals, there is a one-to-one relationship between this sum and the correlation, for the integer scoring of X and Y. That is, the correlation between the variables summarizes the information the data provide about the effect of X on Y.

The likelihood equations equate the sufficient statistics to their expected values. In particular, the likelihood equations for $\{\alpha_j\}$ imply that the fitted marginal counts for Y are the same as the sample marginal counts. This is not the case for cumulative logit and probit models. The log-likelihood function is concave, and the Newton–Raphson iterative method yields the ML estimates of model parameters. The estimators have large-sample normal distributions and their asymptotic SE values are square roots of diagonal elements of the inverse information matrix. For models with these logits, the observed and expected information matrices are the same, so the Fisher scoring fitting algorithm is equivalent to the Newton–Raphson algorihm.

When the data are a nonsparse contingency table, goodness-of-fit tests can use the Pearson or deviance statistics to compare the observed cell counts to the model fitted values. When a model of proportional odds form fits poorly, we can try adding additional terms to the model, such as interactions. Or, we can use the more general model (4.1) or use a model presented in Section 4.3 that nests between the general model and the simple model of proportional odds form.

4.1.5   Example: Opinion on Stem Cell Research and Religious Fundamentalism

Table 4.1 from the 2006 General Social Survey shows the relationship in the United States between opinion about funding stem cell research (Y) and the fundamental-ism/liberalism of one's religious beliefs, stratified by gender. For simplicity, we use scores $x = (1, 2, 3)$ for religious beliefs. For gender $g(1 = \text{females}, 0 = \text{males})$, the model

$$\log \frac{\pi_j}{\pi_{j+1}} = \alpha_j + \beta_1 x + \beta_2 g, \qquad j = 1, 2, 3,$$

describes simultaneously the odds that opinion is "definitely fund" instead of "prob-ably fund," "probably fund" instead of "probably not fund," and "probably not fund" instead of "definitely not fund."

This model is equivalent to the baseline-category logit model

$$\log \frac{\pi_j}{\pi_4} = \alpha_j^* + \beta_1 (4 - j) x + \beta_2 (4 - j) g, \qquad j = 1, 2, 3.$$

The value of the first predictor in this model is set equal to $3x$ in the equation for $\log(\pi_1/\pi_4)$, $2x$ in the equation for $\log(\pi_2/\pi_4)$, and x in the equation for $\log(\pi_3/\pi_4)$. For example, for the liberal category of religious beliefs, the values in the model matrix for the religious beliefs predictor are 9, 6, 3 for each gender, whereas the values for the gender predictor are 3, 2, 1 for females and 0, 0, 0 for males. With

TABLE 4.1. Data on Opinions About Stem Cell Research and Religious Beliefs by Gender, with Conditional Distributions on Stem Cell Research in Parentheses

		Stem Cell Research			
Gender	Religious Beliefs	Definitely Should Fund	Probably Should Fund	Probably Not Fund	Definitely Not Fund
Female	Fundamentalist	34 (22%)	67 (43%)	30 (19%)	25 (16%)
	Moderate	41 (25%)	83 (52%)	23 (14%)	14 (9%)
	Liberal	58 (39%)	63 (43%)	15 (10%)	12 (8%)
Male	Fundamentalist	21 (19%)	52 (46%)	24 (21%)	15 (13%)
	Moderate	30 (27%)	52 (47%)	18 (16%)	11 (10%)
	Liberal	64 (45%)	50 (36%)	16 (11%)	11 (8%)

Source: 2006 General Social Survey.

TABLE 4.2. Output for Fitting Adjacent-Categories Logit Model to Table 4.1 on Funding Stem Cell Research

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept1	1	-0.5001	0.3305	2.29	0.1302
Intercept2	2	0.4508	0.2243	4.04	0.0444
Intercept3	3	-0.1066	0.1647	0.42	0.5178
Religion	4	0.2668	0.0479	31.07	<.0001
Gender	5	-0.0141	0.0767	0.03	0.8539

some software (e.g., PROC CATMOD in SAS; see Table A.2 in the Appendix), we can enter a row of a model matrix for each baseline-category logit at each setting of predictors. Then, after fitting the baseline-category logit model that constrains the effects to be the same for each logit, the estimates of the regression parameters are also the estimates of parameters for the adjacent-categories logit model.

Table 4.2 shows output for this model. For moderates, the estimated odds of opinion "definitely should fund" instead of "probably should fund" are $\exp(0.267) = 1.3$ times the estimated odds for fundamentalists, whereas the estimated odds of opinion "definitely should fund" instead of "definitely should not fund" are $\exp[3(0.267)] = 2.2$ times the estimated odds for fundamentalists (for each gender). For this model, the strongest association results from the extreme categories of each variable. That is, for liberals, the estimated odds of opinion "definitely should fund" instead of "definitely should not fund" are $\exp[2(3)(0.267)] = 5.0$ times the estimated odds for fundamentalists (for each gender). In this sense, the estimated association is relatively strong. Table 4.2 shows that the association is also statistically significant according to a Wald test, with test statistic 31.1 (df = 1) having P-value < 0.0001. The effect of gender is not significant.

The model describes 18 multinomial probabilities (three for each religion × gender combination) using five parameters. The deviance is $G^2 = 12.0$, with df = 13 (P-value = 0.53). This model with a linear trend for the religious beliefs effect and a lack of interaction between it and gender seems adequate.

Similar substantive results occur with a cumulative logit model. Its deviance is $G^2 = 7.5$ with df = 13. The religious beliefs effect is larger ($\hat{\beta}_1 = 0.488$, SE = 0.080), since it refers to the entire response scale rather than only adjacent categories. However, statistical significance is similar, with ($\hat{\beta}_1/$SE) > 5 for each model.

4.1.6 Paired-Category Versus Cumulative Logit Models

For the proportional odds form of model, how can we choose between the adjacent-categories logit form and the cumulative logit form? Since the two types of model tend to fit well in similar situations, the choice cannot usually be based on goodness of fit. One criterion is whether you prefer effects to refer to individual response

categories or instead to groupings of categories using the entire scale or an underlying latent variable. Adjacent-categories logit models describe effects with pairs of individual categories and cumulative logit models describe effects with groupings of categories: namely, cumulative probabilities. In the next section we present yet another possibility, in which effects refer to an individual category relative to a set of categories above that category.

Since effects in cumulative logit models refer to the entire scale, they are usually larger than effects in analogous adjacent-categories logit models. (Recall the discussion in Section 2.2.6 about sizes of effects. Note 2.4 at the end of Chapter 2 showed an approximate relationship, when the effects are weak, between cumulative log odds ratios and local log odds ratios.) The ratio of estimate to standard error, however, is usually similar for the two model types. So one model does not usually have greater power than the other for detecting effects. An advantage of the cumulative logit model is the approximate invariance of effect estimates to the choice and number of response categories, explained in Section 3.3.3. This does not happen with the adjacent-categories logits.

As we discuss in Section 4.3.11, paired-category logit models have the advantage that with retrospective studies (i.e., sampling X at each given value of y), the effects are the same and can still be estimated. Paired-category logit models also have the advantage of being in the exponential family. Hence, reduced sufficient statistics exist, and conditional likelihood methods apply. For example, we can conduct exact small-sample inference for a parameter by eliminating the other "nuisance" parameters from the likelihood function by conditioning on their sufficient statistics.

4.2 CONTINUATION-RATIO LOGIT MODELS

We next present models for *continuation-ratio logits*. There are two types. One set forms the log odds for each category relative to the *higher* categories,

$$\log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_c}, \qquad j = 1, \ldots, c - 1. \tag{4.8}$$

The other set forms the log odds for each category relative to the *lower* categories,

$$\log \frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j}, \qquad j = 1, \ldots, c - 1. \tag{4.9}$$

4.2.1 Logit Models for Sequential Processes

A model using the first type of continuation-ratio logit is useful when a sequential process determines the response outcome. This is the case with duration and development scales, in which a subject passes through each category in order before the response outcome is determined. Examples are survival after receiving a particular medical treatment (<1 year, 1 to 5 years, 5 to 10 years, >10 years), educational

attainment (less than high school, high school, college, postgraduate), and child development through different stages. Let

$$\omega_j = P(Y = j \mid Y \geq j) = \frac{\pi_j}{\pi_j + \cdots + \pi_c}, \qquad j = 1, \ldots, c - 1. \qquad (4.10)$$

The continuation-ratio logits (4.8) are ordinary logits of these conditional probabilities: namely, $\log[\omega_j/(1 - \omega_j)]$. We refer to them as *sequential logits*.

With explanatory variables, continuation-ratio logit models using sequential logits have the form

$$\text{logit } [\omega_j(\mathbf{x})] = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}, \qquad j = 1, \ldots, c - 1. \qquad (4.11)$$

Unlike the cumulative logit model (3.14) having separate effects for each logit, this model provides valid probabilities regardless of predictor values. A simpler model with proportional odds structure is

$$\text{logit } [\omega_j(\mathbf{x})] = \alpha_j + \boldsymbol{\beta}' \mathbf{x}, \qquad j = 1, \ldots, c - 1, \qquad (4.12)$$

in which the effects are the same for each logit (McCullagh and Nelder 1989, p. 164; Tutz 1991). Models with partial proportional odds structure are also possible (Cole and Ananth 2001).

4.2.2 Latent Variable Motivation for Sequential Model

Tutz (1991) provided a latent variable model that induces sequential continuation-ratio logit models. For the model (4.12) of proportional odds form, latent variables $Y_1^*, Y_2^*, \ldots, Y_{c-1}^*$ are assumed to satisfy

$$Y_j^* = \boldsymbol{\beta}' \mathbf{x} + \epsilon_j,$$

where $\{\epsilon_j\}$ are independent from some cumulative distribution function G. Then, for a set of thresholds $\{\alpha_j\}$, there is a process such that the observed categorical outcome

$$Y = 1 \quad \text{if } Y_1^* \leq \alpha_1,$$

and if $Y_1^* > \alpha_1$, then

$$Y = 2, \quad \text{given } Y \geq 2, \quad \text{if } Y_2^* \leq \alpha_2,$$

and so on, with, generally,

$$Y = j, \qquad \text{given } Y \geq j, \quad \text{if } Y_j^* \leq \alpha_j.$$

A transition from category $j - 1$ to category j takes place only if the latent variable that determines the transition is above a threshold that is characteristic of the

category under consideration. The sequential mechanism assumes a binary decision at each step. Only the final resulting category is observable.

This construction leads directly to the model

$$P(Y = j \mid Y \geq j) = G(\alpha_j - \boldsymbol{\beta}'\mathbf{x}).$$

When the underlying distribution for the latent variables is logistic, the link function G^{-1} is the logit link. The resulting model is (4.12), with the sign of each element of $\boldsymbol{\beta}$ changed. An analogous construction with different $\{\boldsymbol{\beta}_j\}$ yields model (4.11). However, as Maddala (1983, p. 51) noted, it is sometimes unrealistic that the random factors that influence responses at the various stages of the sequential process would be independent.

4.2.3 Multinomial Factorization with Sequential Probabilities

For subject i with explanatory variable values $\mathbf{x}_i$, let $\{y_{ij}, \ j = 1, \ldots, c\}$ denote the response indicators. That is, $y_{ij} = 1$ when the response is in category j and $y_{ij} = 0$ otherwise, so $\sum_j y_{ij} = 1$. Let $b(n, y; \omega)$ denote the binomial probability of y successes in n independent trials when the probability of success for each trial is ω, with $b(0, 0; \omega) = 1$. From the expression of the multinomial probability for $y_{i1}, \ldots, y_{ic}$ in the form $p(y_{i1})p(y_{i2} \mid y_{i1}) \cdots p(y_{ic} \mid y_{i1}, \ldots, y_{i,c-1})$, the multinomial mass function for a single observation has factorization

$b[1, \ y_{i1}; \ \omega_1(\mathbf{x}_i)]$

$b[1 - y_{i1}, \ y_{i2}; \ \omega_2(\mathbf{x}_i)] \cdots b[1 - y_{i1} - \cdots - y_{i,c-2}, \ y_{i,c-1}; \ \omega_{c-1}(\mathbf{x}_i)]. \quad (4.13)$

The full likelihood function for all subjects is the product of such multinomial mass functions from the n subjects. The log likelihood is a sum of terms such that different ω_j enter into different terms.

Let $n_1, n_2, \ldots, n_c$ denote the marginal numbers of observations on Y falling in the c response categories. Let S_j denote the set of $n_j^* = n_j + \cdots + n_c$ subjects who make a response in the set of categories $j, \ldots, c$. The likelihood function is a product of $c - 1$ terms. Term j refers to subjects in S_j and looks like the ordinary likelihood function for logistic regression with n_j^* binary outcomes (j vs. $j + 1$ through c combined), such that subject i in group S_j has probability $\omega_j(\mathbf{x}_i)$ of response in category j.

Suppose that parameters in the sequential model specification for logit ω_j are distinct from those for logit ω_k whenever $j \neq k$ [i.e., case (4.11)]. Then a separate set of likelihood equations applies for each sequential binary split for forming the continuation-ratio logits, with separate parameters in each set of equations. Thus, maximizing each set of equations separately maximizes the full log likelihood. That is, separate fitting of models for the various sequential continuation-ratio logits gives the same results as simultaneous fitting. Because these logits refer to a binary response in which one category combines levels of the original scale, separate fitting can use methods and software for binary logistic regression models.

Similarly, overall goodness-of-fit statistics are the sum of goodness-of-fit statistics for the separate fits. For example, consider categorical predictors in model (4.11) and a nonsparse contingency table. The sum of the $c - 1$ separate deviance statistics provides an overall goodness-of-fit statistic pertaining to the simultaneous fitting of $c - 1$ models.

For the proportional odds form of the model (4.12), the likelihood equations from the separate sequential binary splits combine to form a single set of likelihood equations. We discuss this case in Section 4.2.5. Similar remarks about types of models and factorization apply to the other type of continuation-ratio logits, in formula (4.9). However, models with those logits do not give results equivalent to those for models with sequential continuation-ratio logits.

4.2.4 Partitioning Chi-Squared for Independence in Two-Way Tables

A useful application of the multinomial factorization with sequential probabilities relates to testing the hypothesis of independence in a two-way contingency table. Let $\omega_{ij} = P(Y = j \mid Y \geq j, x = i)$. The hypothesis of independence corresponds to the model

$$\text{logit } (\omega_{ij}) = \alpha_j, \qquad j = 1, \ldots, c - 1$$

for $i = 1, \ldots, r$.

First, consider a $2 \times c$ table. The likelihood-ratio statistic G^2 for testing independence, which has df $= c - 1$, partitions into $c - 1$ components. The jth component is G^2 for a 2×2 table where the first column is column j and the second column is columns $j + 1$ through c of the full table. Each component statistic has df $= 1$. For the second type of continuation-ratio logit (4.9), the components are likelihood-ratio statistics for the first two columns, for combining the first two columns and comparing them to the third column, and so on, up to combining the first $c - 1$ columns and comparing them to the last column.

For an $r \times c$ table and a particular type of continuation-ratio logit, each of the $c - 1$ likelihood-ratio statistics relates to a $r \times 2$ table and has df $= r - 1$. More refined partitions contain $(r - 1)(c - 1)$ statistics, each having df $= 1$. See Note 4.3.

4.2.5 Likelihood Equations for Sequential Proportional Odds Model

Let's consider now the proportional odds form (4.12) of the model using the sequential logits. For subject i, the conditional probability $\omega_j(\mathbf{x}_i)$ of response in category j, given response in category j or above, is assumed to satisfy

$$\text{logit } [\omega_j(\mathbf{x}_i)] = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$

in which $\{\beta_k\}$ are the same for each j. We exploit the product binomial factorization (4.13) of the multinomial mass function. For component j, the standard

log-likelihood function for the logistic regression model for the $n_j^* = n_j + \cdots + n_c$ subjects in the group $\mathcal{S}_j$ of subjects who make response in category j or above is[2]

$$L_j = \left(\sum_{\mathcal{S}_j} y_{ij} \right) \alpha_j + \sum_k \left(\sum_{\mathcal{S}_j} y_{ij} x_{ik} \right) \beta_k - \sum_{\mathcal{S}_j} \log \left[1 + \exp \left(\alpha_j + \sum_k \beta_k x_{ik} \right) \right],$$

where $\sum_{\mathcal{S}_j}$ denotes the sum over subjects i in group $\mathcal{S}_j$. The log-likelihood function for the full model is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = L_1 + L_2 + \cdots + L_{c-1}.$$

From the expression for the log-likelihood function, the sufficient statistic for α_j is $\sum_{\mathcal{S}_j} y_{ij} = n_j$. So the total counts in the various response categories for Y are sufficient statistics. The sufficient statistic for β_k is

$$\sum_{\mathcal{S}_1} y_{i1} x_{ik} + \sum_{\mathcal{S}_2} y_{i2} x_{ik} + \cdots + \sum_{\mathcal{S}_{c-1}} y_{i,c-1} x_{ik} = \sum_i (y_{i1} + y_{i2} + \cdots + y_{i,c-1}) x_{ik},$$

where the sum on the right-hand side is over all n subjects. This is the same as the sufficient statistic for β_k for the logistic regression model for the binary split (1 through $c - 1$ combined vs. c).

Differentiating the log-likelihood function with respect to α_j yields the likelihood equation

$$\sum_{\mathcal{S}_j} y_{ij} = \sum_{\mathcal{S}_j} \frac{\exp(\alpha_j + \sum_k \beta_k x_{ik})}{1 + \exp(\alpha_j + \sum_k \beta_k x_{ik})}.$$

The term on the left-hand side is n_j and the term on the right-hand side is $\sum_{\mathcal{S}_j} \omega_j(\mathbf{x}_i)$. This equation implies that the marginal counts for Y equal their fitted values, as in adjacent-categories logit models.

Differentiating the log-likelihood function with respect to β_k yields the likelihood equation

$$\sum_{j=1}^{c-1} \left(\sum_{\mathcal{S}_j} y_{ij} x_{ik} \right) = \sum_{j=1}^{c-1} \left(\sum_{\mathcal{S}_j} \frac{\exp(\alpha_j + \sum_k \beta_k x_{ik}) x_{ik}}{1 + \exp(\alpha_j + \sum_k \beta_k x_{ik})} \right).$$

The term on the right-hand side is

$$\sum_{j=1}^{c-1} \left(\sum_{\mathcal{S}_j} \omega_j(\mathbf{x}_i) x_{ik} \right).$$

The likelihood equation for β_k equates the sufficient statistic for β_k to its expected value. The equation is not quite the same as the equation for β_k for the logistic

[2]For example, see the derivation for equation (5.16) in Agresti (2002).

regression model for the binary split (1 through $c - 1$ combined vs. c), which has the same left-hand side but a different right-hand side. Thus, the ML estimates differ.

We can fit the model using ordinary logistic regression software, by entering a data file that provides the separate binomials that factor to give the multinomial and fitting a model that assumes the same effect for each logit. For example, for a $2 \times c$ table with an indicator x for two groups, consider the model

$$\text{logit} [\omega_j(x)] = \alpha_j + \beta x, \qquad j = 1, \ldots, c - 1.$$

There is one table for each sequential continuation-ratio logit, and each table has the two groups as its rows. The first 2×2 table has response outcome 1 in column 1 and response outcomes 2 to c grouped together in column 2, the second 2×2 table has response outcome 2 in column 1 and response outcomes 3 to c grouped together in column 2, and so on. We can fit the continuation-ratio logit model by fitting to the $c - 1$ stratified 2×2 tables the ordinary logistic regression model having a common treatment effect. See Table A.3 in the Appendix for the following example, and also see Cox (1988).

4.2.6 Example: Tonsil Size and *Streptococcus*

We illustrate continuation-ratio logits using Table 4.3. It cross-classifies a sample of children by their tonsil size and by whether they were carriers of *Streptococcus pyogenes*, a bacterium that is the cause of group A streptococcal infections. The response has three ordered outcomes (not enlarged, enlarged, greatly enlarged). From the conditional distributions shown in Table 4.3, the response distribution is stochastically higher for the carriers. The data have been analyzed by many statisticians, including Tutz (1991), who used continuation-ratio logits with the proportional odds structure.

Tutz (1991) argued that sequential continuation-ratio logits are natural for these data, because of the sequential process by which a subject can develop greatly enlarged tonsils. The tonsils start in the not enlarged state and may become enlarged, perhaps explained by some explanatory variable. If the process continues, the tonsils may become greatly enlarged. The underlying process starts in category 1 (not enlarged) and may transition successively to higher categories until the process

TABLE 4.3. Tonsil Enlargement by Whether a Carrier of Bacteria, with Estimated Conditional Distributions on Tonsil Size in Parentheses

	Tonsil Size		
Carrier	Not Enlarged	Enlarged	Greatly Enlarged
Yes	19 (26%)	29 (40%)	24 (33%)
No	497 (37%)	560 (42%)	269 (20%)

Source: M. Holmes and R. Williams, *J. Hyg. Cambridge*, **52**: 165–179 (1954), with permission.

stops. The latent variable model described in Section 4.2.2 seems plausible. Thus, Tutz used sequential continuation-ratio logits to model (1) the probability π_1 of nonenlarged tonsils, and (2) the conditional probability $\pi_2/(\pi_2 + \pi_3)$ of enlarged tonsils, given that the tonsils were enlarged or greatly enlarged.

Let x indicate whether a child is a carrier of *Streptococcus pyogenes* (1 = yes, 0 = no). The sequential proportional odds model is

$$\log \frac{\pi_1(x)}{\pi_2(x) + \pi_3(x)} = \alpha_1 + \beta x, \qquad \log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta x.$$

To fit the model using ordinary logistic regression model software, we create a data file with the four independent binomials, such as by

```
stratum carrier success failure;
   1       1       19      53
   1       0      497     829
   2       1       29      24
   2       0      560     269
```

Entering the stratum indicator variable in the model, as shown in Table A.3, provides the separate intercept terms for the two logits.

The sample odds ratios for the two strata of binomials to which the continuation-ratio logit model applies are

$$\frac{19 \times 829}{53 \times 497} = 0.598, \quad \frac{29 \times 269}{24 \times 560} = 0.580.$$

These are very similar. In each case, the more desirable outcome is less likely for the carriers of the bacteria. The ML estimate of the carrier effect for the sequential proportional odds model is $\hat{\beta} = -0.5285$ (SE = 0.198), for which $\exp(\hat{\beta}) = 0.59$. For example, given that the tonsils were enlarged, the estimated odds for carriers of having enlarged rather than greatly enlarged tonsils were 0.59 times the estimated odds for noncarriers. The model fits the data very well, with deviance 0.01 (df = 1).

For this model, $\exp(\hat{\beta}) = 0.59$ estimates an assumed common value for a cumulative odds ratio from the first part of the model and a local odds ratio from the second part of the model. By contrast, the cumulative logit model of proportional odds form estimates a common value of $\exp(-0.6025) = 0.55$ for each cumulative odds ratio (model deviance = 0.30, df = 1), and the adjacent-categories logit model of proportional odds form estimates a common value of $\exp(-0.429) = 0.65$ for each local odds ratio (model deviance = 0.24, df = 1). As we would expect, the size of the estimated odds ratio for the continuation-ratio model falls between those for the other two models. According to the deviance, any of these three models is plausible.

The data provide strong evidence of an association. For testing $H_0: \beta = 0$ against $H_a: \beta \neq 0$, the Wald statistic equals $(-0.5285/0.198)^2 = 7.13$ and the likelihood-ratio statistic equals 7.32. The P-values, from the chi-squared distribution with df = 1, are 0.008 and 0.007.

4.2.7 Sequential Models for Grouped Survival Data

In the modeling of survival times with a probability density function f and cumulative distribution function F, the ratio $h(t) = f(t)/[1 - F(t)]$ is called the *hazard function*. Often, survival times are measured with discrete categories, with the response grouped into a set of categories, such as (less than 1 month, 1 month to 1 year, 1 to 3 years, 3 to 5 years, more than 5 years). The ratio $f(t)/[1 - F(t)]$ is analogous to the ratio (4.8) used in sequential continuation-ratio logits. Hence, sometimes continuation-ratio logits are interpreted as log hazards and applied to grouped survival data.

For comparing two groups with grouped survival data, the data are counts in a $2 \times c$ table with the grouped survival times as response categories. For such data we noted at the end of Section 4.2.5 that we can fit the model, logit $\omega_j(x) = \alpha_j + \beta x$, by fitting a standard binary logistic model to a set of $c - 1$ separate 2×2 tables. The hypothesis H_0: $\beta = 0$ of identical response distributions for the two groups is equivalent to the condition that each of these 2×2 tables has a population odds ratio equal to 1.0.

For this application, since the binomials in the separate 2×2 tables are independent, we can apply the Cochran–Mantel–Haenszel test for testing conditional independence in stratified 2×2 tables (e.g., Agresti 2002, Sec. 6.3.2). That test is the score test of H_0: $\beta = 0$ for the model for the stratified tables. In this context, the test is usually referred to as the *logrank test* or the *Mantel–Cox test*. The test can accommodate censored observations, which may occur in some of these 2×2 tables but not others. For Table 4.3 this approach gives a chi-squared test statistic of 7.23 (df $= 1$), similar to the test statistic values given above for Wald and likelihood-ratio tests. The P-value is 0.007 for H_a: $\beta \neq 0$. See Prentice and Gloeckler (1978) for related analyses for grouped survival data and Note 4.2 for related references.

4.3 STEREOTYPE MODEL: MULTIPLICATIVE PAIRED-CATEGORY LOGITS

When a proportional odds model using adjacent-categories logits, continuation-ratio logits, or cumulative logits fits poorly, we can check whether the fit improves by adding other terms, such as interactions. Another approach analyzes whether the fit improves by letting some or all predictors have nonproportional odds form. When all variables have a practical degree of nonproportional odds, with adjacent-categories logits we are left with the general model (4.1) having separate effects for each pair of adjacent categories. But this general model is equivalent to the standard baseline-category logit model,

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}, \qquad j = 1, \dots, c - 1. \tag{4.14}$$

The disadvantage of this model, which treats the response as of nominal scale type, is the lack of parsimony. It has $c - 1$ parameters for each predictor x_k instead of

a single parameter. The number of parameters can be large when either c or the number of predictors is large.

4.3.1 Stereotype Model

Anderson (1984) proposed a paired-category logit model that is nested between the adjacent-categories logit model (4.2) with proportional odds structure and the general model (4.1) for adjacent-category logits and, equivalently, (4.14) for baseline-category logits. For the baseline-category logits, Anderson's model is

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j + \phi_j \boldsymbol{\beta}' \mathbf{x}, \qquad j = 1, \ldots, c - 1. \tag{4.15}$$

Anderson referred to the model as the *stereotype model*. In terms of the response probabilities, the stereotype model is

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \phi_j \boldsymbol{\beta}' \mathbf{x})}{\sum_{k=1}^{c} \exp(\alpha_k + \phi_k \boldsymbol{\beta}' \mathbf{x})}, \qquad j = 1, 2, \ldots, c,$$

with $\alpha_c = \phi_c = 0$.

For logit j, the explanatory variable x_k has coefficient $\phi_j \beta_k$. This represents the log odds ratio for categories j and c of Y with a unit increase in x_k. That is, when $x_k = u + 1$, the odds of response j instead of c are $\exp(\phi_j \beta_k)$ times the odds when $x_k = u$. By contrast, the general baseline-category logit model (4.14) has odds ratio $\exp(\beta_{jk})$ for this effect. That model requires many more parameters for describing all the effects.

The $\{\phi_j\}$ parameters in the stereotype model (4.15) can be regarded as scores for the outcome categories. Since $\phi_j \boldsymbol{\beta} = (\phi_j / C) C \boldsymbol{\beta}$ for any nonzero constant C, these parameters are not identifiable unless we impose a constraint on $\{\phi_j\}$, such as $\phi_1 = 1$. With this constraint, the coefficient β_k of x_k represents the effect of a unit increase in x_k on the log odds of response in category 1 instead of category c.

Given the scores $\{\phi_j\}$, like proportional odds models, the stereotype model has the advantage of requiring only a single parameter to describe the effect of a predictor. So it is more parsimonious than the ordinary baseline-category logit model. We illustrate for the case of four explanatory variables and $c = 3$ outcome categories for Y. The stereotype model for the two baseline-category logits is

$$\log \frac{\pi_1(\mathbf{x})}{\pi_3(\mathbf{x})} = \alpha_1 + \phi_1(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4),$$

$$\log \frac{\pi_2(\mathbf{x})}{\pi_3(\mathbf{x})} = \alpha_2 + \phi_2(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4),$$

with $\phi_1 = 1$. By contrast, the general baseline-category logit model is

$$\log \frac{\pi_1(\mathbf{x})}{\pi_3(\mathbf{x})} = \alpha_1 + \beta_{11} x_1 + \beta_{12} x_2 + \beta_{13} x_3 + \beta_{14} x_4,$$

$$\log \frac{\pi_2(\mathbf{x})}{\pi_3(\mathbf{x})} = \alpha_2 + \beta_{21} x_1 + \beta_{22} x_2 + \beta_{23} x_3 + \beta_{24} x_4.$$

It has three more parameters. The stereotype model achieves the parsimony of a single parameter to describe a predictor effect by using the same scores for each predictor. This may or may not be realistic. Using different scores for each predictor increases flexibility but yields a model equivalent to the general model (4.14).

4.3.2 Stereotype Model for Adjacent-Category Logits

The stereotype model can be expressed in terms of adjacent-categories logits, as

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \nu_j \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1. \tag{4.16}$$

The $\{\nu_j\}$ scores in this model relate to the $\{\phi_j\}$ in the baseline-category logit form of the model (4.15) by

$$\nu_j = \phi_j - \phi_{j+1}, \qquad j = 1, \ldots, c - 1,$$

and

$$\phi_j = \nu_j + \nu_{j+1} + \cdots + \nu_{c-1}, \qquad j = 1, \ldots, c - 1.$$

The discussion at the end of Section 4.1.3 showed how the proportional odds version (4.2) of the adjacent-categories logit model is a special case of a baseline-category logit model with effects $(c - j)\boldsymbol{\beta}$. Therefore, that adjacent-categories logit model is the special case of the stereotype model (4.15) in which $\{\phi_j = c - j\}$, so that $\{\nu_j = 1\}$ in (4.16); that is, the $\{\phi_j\}$ scores are fixed and equally spaced. Equivalently, the scores could be any set of constants that are equally spaced, such as $\{\phi_j = (c - j)/(c - 1)\}$ for the constraints $\phi_1 = 1$ and $\phi_c = 0$ often used with the stereotype model. Thus, if the stereotype model holds and $\{\phi_j\}$ are equally spaced for the baseline-category logits and equivalently, $\{\nu_j\}$ are identical for the adjacent-categories logits, then necessarily the simpler proportional odds adjacent-categories logit model (4.2) holds.

It is often sensible to conduct a likelihood-ratio test comparing the stereotype model with score parameters $\{\phi_j\}$ to the special case with fixed, equally spaced $\{\phi_j\}$, corresponding to the proportional odds version (4.2) of the adjacent-categories logit model. Such a test analyzes whether $\{\phi_j\}$ may depart significantly from being equally spaced. If the simpler model is adequate, it is preferable to use it because of the advantages of model parsimony. When $\{\phi_j\}$ depart significantly from equally spaced but two adjacent scores are similar, it may be sensible to constrain those adjacent scores to be equal and refit the model. This corresponds to collapsing the response scale by combining those two categories. The example in Section 4.3.7 considers these two strategies.

4.3.3 Stereotype Model with Ordered Scores

Nothing inherent in the stereotype model

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j + \phi_j \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1,$$

treats the response variable Y as ordinal. But we've seen that when $\{\phi_j\}$ are a linear function of the category number, the stereotype model is equivalent to an ordinal model: namely, the proportional odds version of the adjacent-categories logit model. Anderson (1984) also proposed an *ordered stereotype model* having the constraint

$$1 = \phi_1 \geq \phi_2 \geq \cdots \geq \phi_c = 0.$$

With such monotone scores, the ordered stereotype model treats Y as ordinal. For a unit increase in a particular predictor x_k, the log odds ratio $\phi_j \beta_k$ for categories j and c of Y is then larger in absolute value when category j is farther from category c.

The baseline-category ordered stereotype model corresponds to the adjacent-categories stereotype model

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + v_j \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1,$$

having the constraint

$$v_j \geq 0, \qquad j = 1, \ldots, c - 1.$$

This implies that the direction of the effect of a predictor x_k is the same for each pair of adjacent categories. For example, a given predictor has either uniformly positive or uniformly negative local log odds ratios with Y.

For the ordered stereotype model, Anderson (1984) noted that the conditional distributions of Y are stochastically ordered according to the values of $\boldsymbol{\beta}'\mathbf{x}$. Specifically, the higher the value of $\boldsymbol{\beta}'\mathbf{x}$, the more the distribution of Y tends to move toward the low end of the response scale. So, for a particular predictor x_k, a value of $\beta_k > 0$ means that the distribution of Y tends to move toward lower values as x_k increases. For the sign of β_k to be such that a positive value means a positive effect, the stereotype models can instead be expressed as

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j - \phi_j \boldsymbol{\beta}'\mathbf{x} \quad \text{or} \quad \log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j - v_j \boldsymbol{\beta}'\mathbf{x}.$$

This parameterization parallels the parameterization logit $P(Y \leq j) = \alpha_j - \boldsymbol{\beta}'\mathbf{x}$ used in cumulative logit models [equation (3.8)] for the same purpose.

In practice, even if $\{\phi_j\}$ are monotone in the stereotype model, pairs of $\{\hat{\phi}_j\}$ in fitting the ordinary model are often out of order because of sampling error. For

example, the standard error of $\hat{\phi}_{j+1} - \hat{\phi}_j$ may be on the same order of size as the difference $\phi_{j+1} - \phi_j$ unless the sample size is quite large. This is not uncommon for pairs of adjacent categories, which have relatively small distances between scores in the ordered stereotype model.

4.3.4 Motivation for Stereotype Form of Model

The stereotype model is an appealing way of obtaining model parsimony, yet permitting a more general model than the proportional odds form of adjacent-categories logit model. But is there any way to motivate this model by a model for underlying latent variables?

Anderson (1984) generalized a construction used to motivate binary logistic regression. Suppose that the conditional distribution of $\mathbf{X}$, given that $Y = j$, is multivariate normal with the same covariance matrix for each j, that is, $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Then, by applying Bayes' theorem, conditional on $\mathbf{x}$, the distribution of Y follows the baseline-category logit model with effects for logit j being

$$\boldsymbol{\beta}'_j = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_c)' \boldsymbol{\Sigma}^{-1}.$$

If the means follow a linear trend for some set of scores, possibly even nonmonotone, then $(\boldsymbol{\mu}_j - \boldsymbol{\mu}_c)' \boldsymbol{\Sigma}^{-1}$ has the form $\phi_j \boldsymbol{\beta}'$ for certain scores. Then the stereotype model holds. If the linear trend has means with the same ordering as the indices, the ordered stereotype model holds.

In proposing the stereotype model as an alternative to standard proportional odds models, Anderson (1984) argued that many ordinal scales are highly subjective and do not result from categorization of a univariate underlying latent variable, but rather, result from a subjective merging of several factors. For example, a physician who diagnoses the severity of a particular illness of a patient, with a scale such as (no illness, mild case, moderate case, severe case), probably takes into account many aspects of a physical examination and available medical tests in applying his or her stereotype of what it means for a patient's condition to be at each category of this scale. Anderson claimed that the stereotype model has greater flexibility than proportional odds models for capturing an inherently multidimensional response, and he proposed even more general stereotype models that are multidimensional.

4.3.5 Interpreting Scores and Checking for Indistinguishability

For the stereotype model (4.15), the log odds ratio comparing values u and $u + 1$ of an explanatory variable x_k in terms of whether the response occurs in category h or j equals

$$\log \frac{P(Y = h \mid \mathbf{x},\ x_k = u + 1) / P(Y = j \mid \mathbf{x},\ x_k = u + 1)}{P(Y = h \mid \mathbf{x},\ x_k = u) / P(Y = j \mid \mathbf{x},\ x_k = u)} = (\phi_h - \phi_j)\beta_k.$$

Hence, the association for these outcome categories is stronger when ϕ_h and ϕ_j are farther apart. Equally spaced $\{\phi_j\}$ implies uniform local conditional association for the various adjacent pairs.

When the stereotype model holds with $\phi_h = \phi_j$, the pair of outcome categories h and j is said to be *indistinguishable*. This means that $\mathbf{x}$ is not predictive between the two categories, in the sense that log odds ratios equal 0 using those two categories and any pair of $\mathbf{x}$ values. Specifically, when $\phi_h = \phi_j$, from (4.15), $\log[\pi_h(\mathbf{x})/\pi_j(\mathbf{x})] = \alpha_h - \alpha_j$ is a constant not dependent on $\mathbf{x}$. In that case, the model still holds with the same $\{\phi_j\}$ if the response scale is collapsed by combining those two categories. Indistinguishability of categories h and j in the general baseline-category logit model (4.14) corresponds to $\boldsymbol{\beta}_h = \boldsymbol{\beta}_j$.

The stereotype model estimates how "close" adjacent response categories j and $j + 1$ are, based on how close $\hat{\phi}_j$ and $\hat{\phi}_{j+1}$ are. This is information not evaluated with cumulative logit models. For those models, the distance between $\{\alpha_j\}$ parameters merely reflects the relative numbers of observations in the various categories.

4.3.6 Stereotype Model for Two-Way Contingency Tables

Let's look at what the stereotype model implies for two-way contingency tables. First, suppose that the single explanatory variable is binary. The model then applies to a $2 \times c$ contingency table. Let $x = 1$ for row 1 and $x = 0$ for row 2. The stereotype model (4.15) then simplifies to

$$\log \frac{\pi_j(x)}{\pi_c(x)} = \alpha_j + \phi_j \beta x, \qquad j = 1, \ldots, c - 1,$$

where $\phi_1 = 1$ for identifiability. The model has $c - 1$ $\{\alpha_j\}$ parameters, $c - 2$ $\{\phi_j\}$ parameters, and the β parameter, for a total of $2(c - 1)$ parameters. This equals the number of multinomial parameters for the two rows ($c - 1$ for each row). Thus, without any restrictions on $\{\phi_j\}$, the model is saturated. The model has the same number of parameters as two multinomial distributions having unrestricted probabilities, so the model perfectly fits any $2 \times c$ contingency table.

For this model, the log odds ratio for outcome categories j and c satisfies

$$\log \frac{\pi_j(1)/\pi_c(1)}{\pi_j(0)/\pi_c(0)} = \phi_j \beta.$$

Consider now the ordered stereotype special case, for which $1 = \phi_1 \geq \phi_2 \geq \cdots \geq \phi_c = 0$. For it, the log odds ratio is monotone in j and the model is no longer saturated. The local log odds ratio for outcome categories j and $j + 1$ equals

$$\log \frac{\pi_j(1)/\pi_{j+1}(1)}{\pi_j(0)/\pi_{j+1}(0)} = (\phi_j - \phi_{j+1})\beta.$$

Thus, for the ordered stereotype model, all the local log odds ratios have the same sign as β. The model then is equivalent to the condition under which there is a uniformly positive association or a uniformly negative association for all the local log odds ratios.

Next suppose that the explanatory variable has r categories. When it is quantitative, or ordinal with fixed scores $\{x_i\}$, the stereotype model is

$$\log \frac{\pi_j(x_i)}{\pi_c(x_i)} = \alpha_j + \phi_j \beta x_i, \qquad j = 1, \ldots, c - 1.$$

When $x_i = i$ represents the row number, the log odds ratio for the 2×2 table consisting of the cells in rows a and b and columns d and e equals

$$\beta(\phi_d - \phi_e)(a - b),$$

where $\phi_c = 0$. The residual df $= (r - 2)(c - 1)$ and the model is unsaturated when $r > 2$. With equally spaced $\{\phi_j\}$, the model then implies uniform local odds ratios.

When the explanatory variable is nominal, a set of $r - 1$ dummy variables can represent its categories, such as $x_i = 1$ for observations from row i and $x_i = 0$ otherwise. The stereotype model is then

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j + \phi_j(\beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1}), \qquad j = 1, \ldots, c - 1.$$

For this model, the log odds ratio for the 2×2 table consisting of the cells in rows a and b and columns d and e equals

$$(\phi_d - \phi_e)(\beta_a - \beta_b).$$

The strength of this association depends on the distance between the scores for the Y categories and the distance between the row effects for the X categories. The residual df $= (r - 2)(c - 2)$ and the model is unsaturated when $r > 2$ and $c > 2$. This model with multiplicative form for the log odds ratios is equivalent to an association model for two-way contingency tables, called the *RC model*, presented in Section 6.5.

4.3.7 Example: Boys' Disturbed Dreams by Age

Anderson (1984) used the stereotype model to analyze Table 4.4, from a study that cross-classified boys by their age and by the severity of their disturbed dreams. Let x_i be the age for row i, using the midpoint scores (6, 8.5, 10.5, 12.5, 14.5). Consider the model

$$\log \frac{\pi_j(x_i)}{\pi_4(x_i)} = \alpha_j + \phi_j \beta x_i, \qquad j = 1, 2, 3,$$

setting $\phi_1 = 0$ and $\phi_4 = 1$. The model has six parameters $(\alpha_1, \alpha_2, \alpha_3, \phi_2, \phi_3, \beta)$ for the 15 multinomial probabilities. The deviance goodness-of-fit statistic is 9.7 (df $= 9$), compared to 32.5 for the independence model (df $= 12$).

TABLE 4.4. Degree of Suffering from Disturbed Dreams, by Age

Age	Degree of Suffering			
	Not Severe (1)	(2)	(3)	Very Severe (4)
5–7	7	4	3	7
8–9	10	15	11	13
10–11	23	9	11	7
12–13	28	9	12	10
14–15	32	5	4	3

Source: A. E. Maxwell, *Analysing Qualitative Data*, Methuen, New York, 1961, p. 70.

Anderson reported estimates for score parameters of

$$\hat{\phi}_1 = 1.0, \quad \hat{\phi}_2 = 0.19 \text{ (SE} = 0.25), \quad \hat{\phi}_3 = 0.36 \text{ (SE} = 0.24), \quad \hat{\phi}_4 = 0.0,$$

and we find $\hat{\beta} = 0.31$. Estimates $\hat{\phi}_2$ and $\hat{\phi}_3$ are out of order, relative to the ordering for the ordered stereotype model. Those two estimates are not far from $\hat{\phi}_4$ relative to their SE values. Anderson also considered the simpler model that constrains

$$\phi_2 = \phi_3 = \phi_4.$$

This simpler model has deviance 11.4 (df $= 11$), 1.7 higher than the model with unconstrained score estimates but with two fewer parameters. The ML estimate of β for this model is identical to the ML estimate of β for the binary logistic regression model

$$\log \frac{\pi_1(x_i)}{\pi_2(x_i) + \pi_3(x_i) + \pi_4(x_i)} = \alpha + \beta x_i.$$

That model has $\hat{\beta} = 0.251$ (SE $= 0.058$), indicating that the probability that a disturbed dream is not severe increases with age.

Alternatively, we consider the special case of the stereotype model with equally spaced $\{\phi_j\}$ severity scores. This is equivalently the adjacent-categories logit model

$$\log \frac{\pi_j(x_i)}{\pi_{j+1}(x_i)} = \alpha_j + \beta x_i, \qquad j = 1, 2, 3,$$

with proportional odds form and a linear effect in the age scores. The deviance is 14.6 (df $= 11$), 4.9 higher (with two fewer parameters) than for the stereotype model with unconstrained score estimates. The model fit has $\hat{\beta} = 0.097$ (SE $= 0.024$). The estimated odds of outcome in the less severe rather than the more severe of two adjacent categories multiplies by $e^{0.097} = 1.10$ for each increase of a year in age. The estimated odds ratio comparing the not severe and very severe categories for a 1-unit increase in age is $\exp[3(0.097)] = 1.34$, compared to $\exp(0.251) = 1.29$ for the constrained model of the preceding paragraph.

By comparison, the cumulative logit model of proportional odds form with a linear trend using the same age scores has deviance 12.4 (df $= 11$). The age effect

is $\hat{\beta} = 0.219$ (SE $= 0.050$). The estimated odds of outcome at the less severe end of the scale multiplies by $e^{0.219} = 1.24$ for each additional year of age. If we do not assume a linear trend for the cumulative logits but, instead, use dummy variables for the categories of age (treating age as nominal scale), the resulting row effects type of model has deviance 7.1 (df $= 8$). The estimated row effects, using constraints that set the final estimate to be 0, are $(-1.82, -1.92, -1.12, -1.14, 0)$. This suggests an alternative way to collapse categories, this time for the age variable using age ranges of 5 to 9 (rows 1 and 2), 10 to 13 (rows 3 and 4), and 14 to 15 (row 5).

4.3.8 ML Fitting of the Stereotype Model

Although the stereotype model has the advantage of being more parsimonious than the ordinary baseline-category logit model with separate effects for each logit, a disadvantage is that it is multiplicative rather than linear in the parameters. That is, the predictor expression has ϕ_j and β multiplied together. Because the predictor is not linear in the parameters, we cannot directly fit the stereotype model with standard software for generalized linear models. Complications also occur in conducting inference for the model parameters, as we discuss in Section 4.3.9.

When $\{\phi_j\}$ are fixed, the model is linear in the parameters. This suggests an iterative two-step approach for fitting nonlinear versions of the model (Goodman 1979a; Greenland 1994): Begin by selecting fixed values for $\{\phi_j\}$. Then estimate β (and $\{\alpha_j\}$) using ordinary ML fitting for baseline-category logit models, by taking predictor k in the model to equal $\phi_j x_k$. If you start with the equally spaced values $\{\phi_j = (c - j)/(c - 1)\}$, this corresponds to the fit of the proportional odds version of the adjacent-categories logit model. At the second step, treating the estimate $\hat{\beta}$ of β from the first stage as fixed and treating $\{\phi_j\}$ as unknown parameters, refit the model to estimate $\{\phi_j\}$. The predictor in the model is now $\hat{\beta}' \mathbf{x}$. This completes the first cycle of the iterative process. In the next cycle you treat the estimates $\{\hat{\phi}_j\}$ from the end of the preceding cycle as fixed and again estimate β, and then treat that estimate of β as fixed to estimate $\{\phi_j\}$. Iterations continue in this way, alternating between a step estimating $\{\phi_j\}$ and a step estimating β, until convergence occurs. This process is not guaranteed to converge to the ML estimates, but it seems to do so when the model fits reasonably well.

A disadvantage of this two-step fitting approach is that the standard errors that software reports at the final iteration for the estimates $\hat{\beta}$ of β are not valid. They treat $\{\phi_j\}$ as fixed, whereas $\{\phi_j\}$ were also estimated in one step of each cycle. Another approach fits the model, recognizing directly both β and $\{\phi_j\}$ as parameters by using an iteratively reweighted least squares algorithm that generates ML estimates (e.g., Holtbrügge and Schumacher 1991).

Various software macros can fit the stereotype model. A useful one is the *gnm* add-on function to R for nonlinear models mentioned in the Appendix.

4.3.9 Inference with the Stereotype Model

The multiplicative nature of the stereotype model makes inference awkward. To illustrate, let's apply the model to $r \times c$ contingency tables. When the predictor is

quantitative or ordinal with row scores $\{x_i\}$, the model is

$$\log \frac{\pi_j(x_i)}{\pi_c(x_i)} = \alpha_j + \phi_j \beta x_i, \qquad j = 1, \ldots, c - 1.$$

The null hypothesis of independence is $H_0: \beta = 0$. When the predictor is nominal with indicator variables $\{z_i\}$ for the rows, the model is

$$\log \frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \alpha_j + \phi_j(\beta_1 z_1 + \cdots + \beta_{r-1} z_{r-1}), \qquad j = 1, \ldots, c - 1.$$

The null hypothesis of independence is $H_0: \beta_1 = \beta_2 = \cdots = \beta_{r-1} = 0$. In both cases, the $\{\phi_j\}$ score parameters are not identifiable under H_0. Because of this, the standard conditions for likelihood-ratio test statistics to have approximate chi-squared distributions are not satisfied. In fact, Haberman (1981) showed that the asymptotic null distribution of the likelihood-ratio statistic for testing independence in the case of the nominal model is the same as the distribution of the largest eigenvalue from a matrix having a Wishart distribution.

For the stereotype model with $\beta \neq 0$, it *is* possible to use ordinary methods to test the indistinguishability hypothesis for a subset of $s < c$ categories, such as $H_0 : \phi_h = \phi_j$ for $s = 2$. The likelihood-ratio statistic has an asymptotic null chi-squared distribution with df $= s - 1$. It is also possible to use a likelihood-ratio test to compare the model to its special case in which the scores are fixed. An example is comparing the stereotype model to the special case with equally spaced scores, which corresponds to the adjacent-categories logit model of proportional odds form. We conducted both these types of inference for the example in Section 4.3.7. It is also possible to compare the model to the general baseline-category logit model (4.14), when that model is unsaturated, to check whether the fit is poorer in using the more parsimonious stereotype model. The following example illustrates.

4.3.10   Example: Back-Pain Prognosis

Anderson (1984) used the stereotype model to analyze a back-pain study with 101 subjects. The response variable was the assessment of back pain after three weeks of treatment using the six ordered categories (worse, same, slight improvement, moderate improvement, marked improvement, complete relief). The three explanatory variables observed at the beginning of the treatment period were $x_1 =$ length of previous attack ($0 =$ short, $1 =$ long), $x_2 =$ pain change (three ordered categories scored $1 =$ getting better, $2 =$ same, $3 =$ worse), and $x_3 =$ lordosis, an inward curvature of a portion of the vertebral column ($0 =$ absent/decreasing, $1 =$ present/increasing). Table 4.5 shows the $2 \times 3 \times 2 \times 6$ contingency table.

The stereotype model for the five baseline-category logits is

$$\log \frac{\pi_j(\mathbf{x})}{\pi_6(\mathbf{x})} = \alpha_j + \phi_j(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3), \qquad j = 1, \ldots, 5.$$

TABLE 4.5. Counts on y = Back Pain by x_1 = Length of Previous Attack, x_2 = Pain Change, and x_3 = Lordosis

$x_1 = 0$		Back Pain[a]						$x_1 = 1$		Back Pain[a]					
x_2	x_3	1	2	3	4	5	6	x_2	x_3	1	2	3	4	5	6
1	0	0	1	0	0	2	4	1	0	0	0	3	0	1	2
1	1	0	0	0	1	3	0	1	1	0	1	0	0	3	0
2	0	0	2	3	0	6	4	2	0	0	3	4	5	6	2
2	1	0	1	0	2	0	1	2	1	1	4	4	3	0	1
3	0	0	0	0	0	2	2	3	0	2	2	1	5	2	0
3	1	0	0	1	1	3	0	3	1	2	0	2	3	0	0

Source: Anderson (1984).
[a]From 1, worse to 6, complete relief.

Table 4.6 shows the ML parameter estimates reported by Anderson. The $\{\hat{\phi}_j\}$ are not monotone, but given the large SE values, it is not implausible that $\{\phi_j\}$ are.

In considering the indistinguishability of categories, Anderson grouped the score parameters into three values, $\phi_1 = 1$, $\phi_2 = \phi_3 = \phi_4$, and $\phi_5 = \phi_6 = 0$, essentially reducing them to a single unknown score parameter that has an ML estimate of 0.30 with SE = 0.13. For this simpler model, Table 4.6 also shows the ML estimates of the effect parameters. Since $\phi_1 = 1$ and $\phi_5 = \phi_6 = 0$, exponentiating a $\hat{\beta}_k$ value gives an estimated odds ratio for the odds of response "worse" instead of "marked improvement" or "complete relief." For example, for lordosis present or increasing and fixed values of length of previous attack and pain change, the estimated odds of the response "worse" instead of "marked improvement" or "complete relief" were exp(1.05) = 2.86 times the estimated odds for lordosis absent or decreasing. The estimate is very imprecise, as the corresponding Wald 95% confidence interval is exp[1.05 ± 1.96(0.47)], or (1.14, 7.18). Since $\hat{\phi}_2 = \hat{\phi}_3 = \hat{\phi}_4 = 0.30$ and $\phi_5 = \phi_6 = 0$, exp($0.30\hat{\beta}_k$) is an estimated odds ratio for the odds of the response "same," "slight improvement," or "moderate improvement" instead of the response "marked improvement" or "complete relief." For example, for lordosis present or increasing and fixed values of the length of the previous attack and pain change, the estimated odds of the response "same," "slight improvement," or "moderate improvement" instead of the response "marked improvement"

TABLE 4.6. ML Estimates for Stereotype Models Fitted to Back-Pain Data of Table 4.5

Ordinary model	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$	$\hat{\phi}_5$	$\hat{\phi}_6$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Estimate	1	0.31	0.35	0.51	0.14	0	2.63	2.15	1.31
SE	—	0.13	0.14	0.17	0.10	—	0.93	0.75	0.51
Simpler model[a]									
Estimate	1	0.30	0.30	0.30	0	0	2.79	1.80	1.05
SE	—	0.13	0.13	0.13	—	—	1.31	0.74	0.47

[a]Sets $\phi_2 = \phi_3 = \phi_4$ and $\phi_5 = \phi_6$.

or "complete relief" were $\exp[0.30(1.05)] = 1.37$ times the estimated odds for lordosis absent or decreasing.

Anderson and Phillips (1981) also fitted the cumulative logit model of proportional odds form

$$\log \frac{\pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \cdots + \pi_6(\mathbf{x})} = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \qquad j = 1, \ldots, 5.$$

The effect estimates are $\hat{\beta}_1 = 1.515$ (SE = 0.402), $\hat{\beta}_2 = 0.486$ (SE = 0.265), and $\hat{\beta}_3 = 0.866$ (SE = 0.374), with SE values based on the observed information matrix. According to this fit, for lordosis present or increasing and fixed values of the length of the previous attack and pain change, the estimated odds of the response "worse" instead of "same" or "improvement" or "complete relief" were $\exp(0.866) = 2.38$ times the estimated odds for lordosis absent or decreasing. The model has one fewer parameter than the simpler stereotype model (with $\phi_2 = \phi_3 = \phi_4$ and $\phi_5 = \phi_6$). According to a fit criterion such as AIC, the simpler stereotype model is preferred to this cumulative logit model. The maximized log-likelihood values are -159.0 for the cumulative logit model and -154.4 for the simpler stereotype model having only one additional parameter. However, this comparison must take into account that the simpler stereotype model was suggested by the fit of the ordinary stereotype model. That stereotype model is also preferred to the cumulative logit model according to AIC, as its maximized log likelihood is -151.6 with four more parameters (the difference being larger than the number of parameters).

Another possible model is the simpler stereotype model with equally spaced $\{\phi_j\}$, which is equivalent to the adjacent-categories logit model of proportional odds form

$$\log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

The quality of fit is similar to the cumulative logit model, with a maximized log likelihood of -160.1. The effect estimates are $\hat{\beta}_1 = 0.605$ (SE = 0.180), $\hat{\beta}_2 = 0.217$ (SE = 0.116), and $\hat{\beta}_3 = 0.320$ (SE = 0.162). The estimates and SE values are on the order of 40% of the size of those for the cumulative logit model. According to this fit, for lordosis present or increasing and fixed values of the length of the previous attack and pain change, the estimated odds of response in the worse instead of the better of two adjacent categories were $\exp(0.320) = 1.38$ times the estimated odds for lordosis absent or decreasing.

Any of these models are much more parsimonious than a full baseline-category logit model having separate parameters for each logit (i.e., treating the response as nominal). Such a model has four parameters for each logit, a total of 20 parameters, compared to eight parameters for the cumulative logit or adjacent-categories logit model of proportional odds form (five α_j and three β_k), nine parameters for the simpler stereotype model, and 12 parameters for the ordinary stereotype model. The full baseline-category logit model does not give a significantly better fit than the

two stereotype models, as its maximized log likelihood of -149.5 is only slightly higher.

4.3.11 Using Paired-Category Logit Models with Retrospective Studies

Some studies, such as retrospective studies in epidemiology, sample $\mathbf{X}$ conditional on y instead of Y conditional on $\mathbf{x}$. Such studies take subjects with certain values on y (such as having stage II of a certain cancer, stage I of the cancer, or no disease) and then observe $\mathbf{x}$ values that measure subject characteristics such as past smoking behavior. The sampling is then *outcome-dependent sampling* instead of independent multinomial sampling on Y. With outcome-dependent sampling, the effects are preserved and can be estimated for paired-category logit models, such as adjacent-categories logit models and stereotype models. That is, we can use the same estimates and SE values as we would obtain by treating the data as ordinary independent multinomial observations on Y.

To illustrate, for the stereotype model,

$$\frac{\pi_j(\mathbf{x})}{\pi_c(\mathbf{x})} = \exp(\alpha_j + \phi_j \boldsymbol{\beta}'\mathbf{x}), \qquad j = 1, \ldots, c-1,$$

suppose that the sampling fractions from the various categories of y are $\{f_1, f_2, \ldots, f_c\}$. For the sampled population, the odds in terms of pairs of categories of Y are

$$\frac{P(Y = j \mid \mathbf{x}, \text{ sampled})}{P(Y = c \mid \mathbf{x}, \text{ sampled})} = \frac{f_j \exp(\alpha_j + \phi_j \boldsymbol{\beta}'\mathbf{x}) / \sum_k \exp(\alpha_k + \phi_k \boldsymbol{\beta}'\mathbf{x})}{f_c / \sum_k \exp(\alpha_k + \phi_k \boldsymbol{\beta}'\mathbf{x})}$$

$$= \exp(\alpha_j^* + \phi_j \boldsymbol{\beta}'\mathbf{x}),$$

where $\alpha_j^* = \alpha_j + \log(f_j/f_c)$. The parameters for the effects of the predictors are preserved. Those effect parameters can be estimated consistently with outcome-dependent data (but the intercept terms cannot be). This is not the case with models, such as cumulative logit models, that group outcome categories together. Mukherjee and Liu (2008) presented necessary and sufficient conditions for the link functions that allow for the equivalence of prospective and retrospective inference for multinomial models. They showed that the equivalence does not hold beyond paired-category logit models. For related work, see Greenland (1994) and Mukherjee et al. (2007, 2008).

CHAPTER NOTES

Section 4.1: Adjacent-Categories Logit Models

4.1. Adjacent-categories logit models or models equivalent to them have been presented by Simon (1974), Andrich (1978, 1979), Goodman (1979a, 1983, 1991), Masters (1982), and Magidson (1996). See also Agresti (1992b) and Böckenholt and Dillon (1997) for modeling paired comparison data, Hirji (1992) for exact

small-sample inference, Lipsitz et al. (1996) for a test of fit patterned after the Hosmer–Lemeshow test for binary logistic regression, Sobel (1997) for specialized structure when there is a middle category, Hartzel et al. (2001a,b) for random effects models, and Agresti and Lang (1993b) and DeSantis et al. (2008) for latent class models.

Section 4.2: Continuation-Ratio Logit Models

4.2. Thompson (1977) used continuation-ratio logits in modeling discrete survival-time data. When the lengths of time intervals approach zero, his model converges to the Cox proportional hazards model. See also Section 5.3.3, Prentice and Gloeckler (1978), Fienberg and Mason (1979), Aranda-Ordaz (1983), Berridge and Whitehead (1991), Ten Have and Uttal (1994), Heagerty and Zeger (2000a), Hedeker et al. (2000), Albert and Chib (2001) for a Bayesian approach, Fahrmeir and Tutz (2001, Chap. 6.5), Tutz and Binder (2004), and Grilli (2005). For other applications of continuation-ratio logits, see Fienberg (1980, pp. 114–116), Cox and Chuang (1984), Läärä and Matthews (1985), Cox (1988), Armstrong and Sloan (1989), Tutz (1989, 1990, 1991), Berridge and Whitehead (1991), Ryan (1992), Barnhart and Sampson (1994), Greenland (1994), Joffe and Greenland (1995), Smith et al. (1996), Yee and Wild (1996), Lindsey et al. (1997), Scott et al. (1997), Coull and Agresti (2000), Guisan and Harrell (2000), Dos Santos and Berridge (2000), Kvist et al. (2000), Ten Have et al. (2000), Hemker et al. (2001), and Fu and Simpson (2002).

4.3. The continuation odds ratios defined in Section 2.2.4 apply to a set of 2×2 tables for which the likelihood-ratio (LR) statistic for testing independence in the $r \times c$ table partitions exactly into a sum of $(r - 1)(c - 1)$ components. Each component is the LR statistic computed for a 2×2 table. The $(r - 1)(c - 1)$ separate 2×2 tables are

$$
\begin{array}{c|c}
n_{ij} & \sum_{b > j} n_{ib} \\
\hline
\sum_{a > i} n_{aj} & \sum_{a > i} \sum_{b > j} n_{ab}
\end{array}
$$

for $i = 1, \ldots, r - 1$ and $j = 1, \ldots, c - 1$ (Lancaster 1949). Such partitionings do not apply to the other ordinal odds ratios presented in Section 2.2.

Section 4.3: Stereotype Model: Multiplicative Paired-Category Logits

4.4. The stereotype model has been discussed by Anderson (1984), Greenwood and Farewell (1988), DiPrete (1990), Holtbrügge and Schumacher (1991), Greenland (1994), Joffe and Greenland (1995), Ananth and Kleinbaum (1997), Guisan and Harrell (2000), Lall et al. (2002), and Kuss (2006). See also references for the related multiplicative RC model for two-way contingency tables in Section 6.5.2 and Note 6.7. Cox and Chuang (1984) proposed similar multiplicative logit models for contingency tables using baseline-category logits and cumulative logits.

Greenland (1994) described underlying processes for which he believed stereotype models are more natural than cumulative logit models. Johnson (2007) used it for discrete choice modeling of an ordinal response. Greenland (1994) suggested using the bootstrap to estimate valid standard errors. Kuss (2006) used PROC NLMIXED in SAS to fit the model with a quasi-Newton maximization method that uses finite-difference methods for the first derivatives. This method gradually builds up an approximation to the matrix of second partial derivatives as the iterations proceed, and at convergence its inverse gives a valid estimate for the asymptotic covariance matrix. Yee and Hastie (2003) suggested another approach.

EXERCISES

4.1. For the *row effects* model (4.4), show that the sufficient statistics for $\{\tau_i\}$ are sample means computed within the rows, using the column scores $(1, 2, \ldots, c)$.

4.2. If cumulative logit and adjacent-categories logit models with proportional odds structure both fit a data set well, explain why the parameter estimates from the cumulative logit model would probably be larger. Does this imply that it is easier for effects to achieve statistical significance with that model? Explain.

4.3. Prove factorization (4.13) for the multinomial distribution.

4.4. Summarize some advantages and disadvantages of the stereotype model compared to the ordinary multinomial logit model using baseline-category logits.

4.5. Analyze Table 4.1 with (a) a cumulative logit model and (b) a continuation-ratio logit model. Compare and contrast results to those obtained in Section 4.1.5 using adjacent-categories logit models.

4.6. Refer to Exercise 2.7. Analyze these data using methods of this chapter.

CHAPTER 5

Other Ordinal Multinomial Response Models

In Chapters 3 and 4 we presented several multinomial models for ordinal response variables that use the logit link function. Among these, most commonly used are the models for logits of cumulative probabilities in Chapter 3 that have proportional odds structure.

For binary data, models can use link functions other than the logit. Best known of these is the *probit model*. Similarly, in the ordinal response case, other link functions are possible. In Section 5.1 we present a general model with a family of link functions for cumulative probabilities. Sections 5.2 and 5.3 cover two important special cases, with the probit link function and a log-log link function. In Section 5.4 we present a further generalization to allow explanatory variables to have dispersion effects as well as location effects. In Section 5.5 we show how to apply such models to construct receiver operating characteristic (ROC) curves as a way of assessing diagnostic tests.

Models using nonlinear link functions, such as the logit and probit, have the disadvantage that they can be difficult for nonstatisticians to understand and to interpret. In Section 5.6 we consider a simpler model for ordinal responses that assigns scores to the outcome categories and models the mean response directly as a linear function of the predictors.

5.1 CUMULATIVE LINK MODELS

Let h denote an arbitrary link function. The model

$$h[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1, \tag{5.1}$$

links the cumulative probabilities to a linear predictor. As in the proportional odds model (3.6), the effects of $\mathbf{x}$ in (5.1) are the same for each cumulative

probability, $j = 1, \ldots, c - 1$. We refer to this class of models as *cumulative link models*.

In Section 3.3.2 we showed that the homogeneous effects assumption holds when a linear regression holds for a continuous latent variable Y^*. Specifically, model (5.1) with $-\boldsymbol{\beta}$ rather than $+\boldsymbol{\beta}$ in the linear predictor results when Y is a discrete measurement of a latent variable Y^* that satisfies the regression model

$$Y^* = \boldsymbol{\beta}'\mathbf{x} + \epsilon,$$

with ϵ having a cdf of some standard form G. The link function h relates to G by $h(u) = G^{-1}(u)$, that is, the inverse of the continuous cdf G. For example, the logit link function $h(u) = \log[u/(1-u)]$ is the inverse of the standard logistic cdf. Assuming that Y^* has, conditional on $\mathbf{x}$, a logistic distribution with constant variance leads to the cumulative logit model of proportional odds form. The parameters $\{\alpha_j\}$ are category cutpoints on a standardized version of the latent scale. In this sense, cumulative link models are regression models, using a linear predictor $\boldsymbol{\beta}'\mathbf{x}$ to describe effects of explanatory variables on ordinal categorical measurement Y of Y^*. Using $+\boldsymbol{\beta}$ rather than $-\boldsymbol{\beta}$ in the linear predictor simply results in a change of sign of $\hat{\boldsymbol{\beta}}$.

5.1.1   Common Link Functions for Cumulative Link Models

After the logit, the most commonly used link function for cumulative link models is the *probit*, which is the inverse of the standard normal cdf. Cumulative link models using the probit link function are called *cumulative probit models*. We study this model in Section 5.2. It results directly when the latent variable model is the standard regression model for which the conditional distribution of Y^*, given the predictors, is normal with constant variance.

Another useful link function is the *complementary log-log* link, $\log\{-\log[1 - P(Y \leq j)]\}$, and the related *log-log* link, $\log\{-\log[P(Y \leq j)]\}$. Unlike the logit and probit links, these link functions are not symmetric. With a continuous predictor x, for example, $P(Y \leq j)$ approaches 0 at a different rate than it approaches 1. These cumulative link models are presented in Section 5.3.

Another symmetric link function, less commonly used, is the inverse of a Cauchy cdf, sometimes called the *cauchit* link. The Cauchy distribution has much thicker tails than the normal or the logistic. So this link function is appropriate when the conditional distributions of an underlying latent variable have a substantial chance of extreme values for y^*, in either direction. This link function is $\tan\{\pi[P(Y \leq j) - 0.5]\}$ for the mathematical constant π.

5.1.2   ML Estimation for Cumulative Link Models

For subject i, let $y_{ij} = 1$ if $y_i = j$ and let $y_{ij} = 0$ otherwise, $i = 1, \ldots, n$. Then $E(Y_{ij}) = \pi_j(\mathbf{x}_i)$, the probability that observation i with explanatory variable values $\mathbf{x}_i$ falls in category j. McCullagh (1980) and Fahrmeir and Tutz (2001, pp. 76,

88–89) treated cumulative link models as multivariate GLMs: The multivariate distribution is the multinomial, and the link function h applies to a vector of means $(\pi_1(\mathbf{x}_i), \ldots, \pi_c(\mathbf{x}_i))$. As we explained in Section 3.2, we can also view this in terms of a composite link function.

Let $G = h^{-1}$ denote the inverse link function for the cumulative link model, such as the standard normal cdf for the cumulative probit model. With independent observations, we obtain the likelihood function by substituting $G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)$ for $P(Y \leq j \mid \mathbf{x}_i)$ in the product of multinomial probability mass functions,

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{c}\pi_j(\mathbf{x}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{c}\left[P(Y_i \leq j \mid \mathbf{x}_i) - P(Y_i \leq j-1 \mid \mathbf{x}_i)\right]^{y_{ij}}\right\}.$$

The log-likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log[G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - G(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)].$$

Let g denote the derivative of G, that is, the probability density function corresponding to the cdf G, and let δ_{jk} denote the Kronecker delta, $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. Then the likelihood equations are

$$\frac{\partial L}{\partial \beta_k} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{g(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - g(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - G(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} = 0,$$

$$\frac{\partial L}{\partial \alpha_k} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\frac{\delta_{jk}g(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - \delta_{j-1,k}g(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - G(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} = 0.$$

McKelvey and Zavoina (1975) derived the information matrix for the cumulative probit model. Substituting G in place of their standard normal cdf yields this matrix for the general cumulative link model. Denote $z_{ij} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i$. Then the second partial derivatives are

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_\ell} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}x_{i\ell}\left\{\frac{[G(z_{ij}) - G(z_{i,j-1})][g(z_{i,j-1})z_{i,j-1} - g(z_{ij})z_{ij}]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right.$$

$$\left. - \frac{[g(z_{i,j-1}) - g(z_{ij})]^2}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right\},$$

$$\frac{\partial^2 L}{\partial \beta_k \partial \alpha_\ell} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}\left\{\frac{[g(z_{i,j-1}) - g(z_{ij})][g(z_{ij})\delta_{j\ell} - g(z_{i,j-1})\delta_{j-1,\ell}]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right.$$

$$\left. - \frac{[G(z_{ij}) - G(z_{i,j-1})][g(z_{ij})z_{ij}\delta_{j\ell} - g(z_{i,j-1})z_{i,j-1}\delta_{j-1,\ell}]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right\},$$

$$\frac{\partial^2 L}{\partial \alpha_k \, \partial \alpha_\ell} = \sum_{i=1}^{n} \sum_{j=1}^{c}$$

$$y_{ij} \left\{ \frac{[G(z_{ij}) - G(z_{i,j-1})][g(z_{i,j-1})z_{i,j-1}\delta_{j-1,k}\delta_{j-1,\ell} - g(z_{ij})z_{ij}\delta_{jk}\delta_{j\ell}]}{[G(z_{ij}) - G(z_{i,j-1})]^2} \right.$$

$$\left. - \frac{[g(z_{ij})\delta_{jk} - g(z_{i,j-1})\delta_{j-1,k}][g(z_{ij})\delta_{j\ell} - g(z_{i,j-1}\delta_{j-1,\ell}]}{[G(z_{ij}) - G(z_{i,j-1})]^2} \right\}.$$

Replacing y_{ij} in these terms by $E(Y_{ij}) = [G(z_{ij}) - G(z_{i,j-1})]$ (i.e., the probability of category j for subject i) and taking negatives yields the expected information matrix. The inverse of this matrix, with parameters replaced by their ML estimates, is the estimated asymptotic covariance matrix of the parameter estimates.

Unlike models using paired-category logits such as adjacent-categories logit models, the model does not have reduced sufficient statistics. Similarly, the likelihood equations do not have the simple form of equating sufficient statistics to their expected values. For example, unlike adjacent-categories logit models, cumulative link models need not have fitted marginal counts for Y that are the same as the sample marginal counts. McCullagh presented a Fisher scoring algorithm for ML estimation. A unique maximum of the likelihood function occurs with sufficiently large n. Burridge (1981) and Pratt (1981) showed that the log-likelihood function is concave for many cumulative link models, including the models with logit, probit, and complementary log-log link functions. Because of the concavity, iterative algorithms converge rapidly to the ML estimates unless any estimates are infinite or do not exist. Remarks in Section 3.4.5 about infinite estimates for cumulative logit models also apply to the corresponding cumulative link models.

5.1.3 Interpreting Effects on an Underlying Latent Response

The interpretation of the effects β in a cumulative link model depends on the link function h. For the logit link, for instance, Section 3.3.1 showed that β_k is the effect of a unit increase in x_k on the log odds for each cumulative probability, controlling for the other predictors; thus, $\exp(\beta_k)$ is a cumulative odds ratio using any collapsing of the ordinal response for values of x_k that differ by 1.

Regardless of the link function, an alternative interpretation refers to the underlying latent variable model, for which

$$Y^* = \beta' \mathbf{x} + \epsilon,$$

where ϵ has a cdf of some standard form G. A unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k, keeping the other predictor values fixed. The size of the effect is relative to the spread of the conditional distribution of Y^*, which depends on the standard deviation of the cdf G. When Y^* is scaled such that ϵ has standard deviation σ, a 1-unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k/σ standard deviations of the conditional distribution of Y^*. Common values are

$\sigma = \pi/\sqrt{3}$ for the standard logistic distribution for ϵ and $\sigma = 1$ for the standard normal distribution.

Alternatively, standardized effects can be expressed as multiples of the unconditional standard deviation of Y^*, as is often done in ordinary regression. Assuming that ϵ and the explanatory variables $\mathbf{X}$ are uncorrelated, the unconditional variance of Y^* is

$$\boldsymbol{\beta}'\text{Var}(\mathbf{X})\boldsymbol{\beta} + \text{Var}(\epsilon),$$

where $\text{Var}(\mathbf{X})$ denotes the covariance matrix of $\mathbf{X}$. So the standardized coefficient is (Winship and Mare 1984)

$$\beta_k^* = \frac{\beta_k}{\sqrt{\boldsymbol{\beta}'\text{Var}(\mathbf{X})\boldsymbol{\beta} + \text{Var}(\epsilon)}}.$$

Multiplying this standardized coefficient by the standard deviation of X_k gives a fully standardized coefficient for which effects refer to a standard deviation change in X_k. Such fully standardized coefficients can be useful for comparing effects of predictors having different units of measurement.

5.2 CUMULATIVE PROBIT MODELS

Denote the cdf of the standard normal distribution by Φ. This has an appearance very similar to the symmetric S-shape of the cdf for the logistic distribution with mean 0 and standard deviation 1. The *cumulative probit model* is

$$\Phi^{-1}[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \qquad j = 1, \ldots, c - 1. \tag{5.2}$$

Some fields call it the *ordered probit model*. As in the proportional odds model with the logit link, the effect $\boldsymbol{\beta}$ is the same for each cumulative probability. But it is not appropriate to call this model a "proportional odds" model because probit model interpretations do not apply to odds or to odds ratios.

The cumulative probit model describes the cumulative probabilities directly as

$$P(Y \leq j) = \Phi(\alpha_j + \boldsymbol{\beta}'\mathbf{x}), \qquad j = 1, \ldots, c - 1.$$

For example, $P(Y \leq j \mid \mathbf{x}) = \frac{1}{2}$ for $\mathbf{x}$ values such that $\alpha_j + \boldsymbol{\beta}'\mathbf{x} = 0$, since $\Phi(0) = \frac{1}{2}$ is the probability that a standard normal random variable falls below 0. Similarly, since the central 68% of a standard normal distribution falls between -1 and 1, $P(Y \leq j \mid \mathbf{x}) = 0.16$ for $\mathbf{x}$ values such that $\alpha_j + \boldsymbol{\beta}'\mathbf{x} = -1$, and $P(Y \leq j \mid \mathbf{x}) = 0.84$ for $\mathbf{x}$ values such that $\alpha_j + \boldsymbol{\beta}'\mathbf{x} = 1$.

In some fields that place strong emphasis on latent variable models, particularly econometrics, the cumulative probit model is more popular than the cumulative logit model. McKelvey and Zavoina (1975) gave expressions for the information matrix for the model.

5.2.1 Interpreting Parameters in Cumulative Probit Models

The cumulative probit model (5.2) generalizes the binary probit model to ordinal responses. It is implied by a model in which an underlying continuous latent variable Y^* satisfies an ordinary regression model $Y^* = \beta' x + \epsilon$ in which ϵ has a normal distribution with mean 0 and constant standard deviation. The observed ordinal scale provides no information about variability for the underlying latent variable. So without loss of generality we can let the standard deviation of ϵ be 1. (Recall that this latent variable model actually generates the cumulative link model with linear predictor $\alpha_j - \beta' x$ rather than $\alpha_j + \beta' x$.)

How can we interpret parameters in terms of a latent variable Y^*? Having the inverse standard normal cdf as the link function corresponds to a standard deviation for ϵ that equals 1. This is also the conditional standard deviation for Y^*. So for coefficient β_k of x_k, a unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k conditional standard deviations of Y^*, keeping the other predictor values fixed.

5.2.2 Comparison of Cumulative Logit and Cumulative Probit Models

Because logistic and normal cdf's having the same mean and the same standard deviation look so similar, cumulative probit models and the corresponding cumulative logit models fit well in similar situations. Whereas the standard normal distribution has mean 0 and standard deviation 1, however, the standard logistic distribution has mean 0 and standard deviation $\pi/\sqrt{3} = 1.81$. Because of this, their ML estimates are not on the same scale. The standard normal cdf at a point z is well approximated by the standard logistic cdf at the point $(15\pi/16\sqrt{3})z = 1.7z$. Typically, ML estimates from cumulative logit models are about 1.6 to 1.8 times the ML estimates from cumulative probit models.

The coefficient β_k of x_k in the cumulative logit model with linear predictor $\alpha_j - \beta' x$ has the interpretation that a unit increase in x_k corresponds to an increase in $E(Y^*)$ of β_k, keeping the other predictor values fixed, when Y^* has standard deviation $\pi/\sqrt{3}$. Thus, a unit increase in x_k corresponds to an increase of $\beta_k/(\pi/\sqrt{3})$ standard deviations in the underlying response scale. For example, if $\hat{\beta} = 0.345$ with a single quantitative predictor, as in the following example, a 1-unit increase in x corresponds to an increase of $0.345/(\pi/\sqrt{3}) = 0.19$ conditional standard deviation in the mean of the underlying latent response.

5.2.3 Example: Religious Fundamentalism by Educational Degree

Table 5.1 cross-classifies subjects by their highest educational degree and by whether they are fundamentalist, moderate, or liberal in their religious beliefs. The table contains data from every GSS since 1972. (We use all the years in order to show some effects of having a very large sample size, $n = 49,040$.)

Consider the cumulative link model with link function h,

$$h[P(Y \leq j)] = \alpha_j + \beta x_i, \qquad j = 1, 2,$$

TABLE 5.1. Data on Highest Educational Degree and Religious Beliefs, with Conditional Distributions on Religious Beliefs in Parentheses

Highest Degree	Religious Beliefs		
	Fundamentalist	Moderate	Liberal
Less than high school	4,913 (43%)	4,684 (41%)	1,905 (17%)
High school	8,189 (32%)	11,196 (44%)	6,045 (24%)
Junior college	728 (29%)	1,072 (43%)	679 (27%)
Bachelor	1,304 (20%)	2,800 (43%)	2,468 (38%)
Graduate	495 (16%)	1,193 (39%)	1,369 (45%)

Source: General Social Survey.

using scores $\{x_i\}$ for the rows to treat educational degree in a quantitative manner. Using the row numbers as the scores, $\hat{\beta} = -0.206$ (SE = 0.0045) with the probit link and -0.345(SE = 0.0075) with the logit link. From the logit model estimate, for each of the two cutpoints for the response variable, the estimated odds of response in the fundamentalist rather than the liberal direction multiply by $\exp(-0.345) = 0.71$ for each category increase in highest degree. So the estimated cumulative odds ratio for comparing those with a graduate degree to those with less than a high school degree is $\exp[4(-0.345)] = 0.25$. For example, the estimated odds of response fundamentalist rather than moderate or liberal for those with less than a high school education are $1/0.25 = 4.0$ times the estimated odds for those with a graduate degree.

Next we consider effects in terms of an underlying continuous latent variable for religious beliefs, with higher values corresponding to responses that are more liberal. The $\hat{\beta}$ values for the logit and probit links with the $\alpha_j - \beta x_i$ parameterization are positive. From the probit model estimate of $\hat{\beta} = 0.206$, for each category increase in highest degree, the mean of the latent response on religious beliefs is estimated to increase by about 0.21 conditional standard deviation of that underlying scale. Similarly, since $\hat{\beta} = 0.345$ for the logit link, for each category increase in highest degree, the mean of the latent response on religious beliefs is estimated to increase by about $0.345/(\pi/\sqrt{3}) = 0.19$ conditional standard deviation of that underlying scale.

For the unit-spaced scores for highest degree, the estimated standard deviation of that explanatory variable is $s_x = 1.144$. The estimated unconditional standard deviation of the latent response variable is

$$\sqrt{(\hat{\beta}s_x)^2 + \text{Var}(\epsilon)}.$$

This equals 1.11 for the probit link and 1.92 for the logit link. So the standardized effects of highest degree in terms of the unconditional variability of the latent response are $(0.206)/1.11 = 0.185$ for the probit link and $(0.345)/1.92 = 0.180$ for the logit link. The estimated effect is similar for the two models.

According to formal goodness-of-fit tests, both models show lack of fit. The deviance is 48.7 for the cumulative probit model and 45.4 for the cumulative

logit model (df = 7 for each, $P < 0.0001$). With such an enormous sample size, however, we expect a test of nearly any hypothesis to be statistically significant. We address below whether the lack of fit is also practically significant.

The more general model that has row effects rather than a linear trend for the effect of educational degree,

$$h[P(Y \leq j)] = \alpha_j + \tau_i, \qquad j = 1, 2,$$

treats educational degree as a qualitative factor. Even with such a large n, this model fits adequately according to goodness-of-fit tests, with deviance 5.2 for the cumulative probit model and 2.4 for the cumulative logit model (df = 4). With the constraint $\tau_5 = 0$, the ML estimates of the row effects are

$$\hat{\tau}_1 = 0.83, \quad \hat{\tau}_2 = 0.56, \quad \hat{\tau}_3 = 0.46, \quad \hat{\tau}_4 = 0.17, \quad \hat{\tau}_5 = 0$$

for the cumulative probit model, and

$$\hat{\tau}_1 = 1.39, \quad \hat{\tau}_2 = 0.94, \quad \hat{\tau}_3 = 0.78, \quad \hat{\tau}_4 = 0.29, \quad \hat{\tau}_5 = 0$$

for the cumulative logit model. For example, the estimated odds of response fundamentalist rather than moderate or liberal for those with less than a high school education are $\exp(1.39) = 4.0$ times the estimated odds for those with a graduate degree. From the probit estimates, the mean of the underlying latent response on religious beliefs is estimated to be about 0.83 (conditional) standard deviation of that underlying scale lower (i.e., more fundamentalist) for those with less than a high school education than for those with a graduate degree.

For the row effects model and the linear trend model, the estimates provide similar information. The estimates for the cumulative logit model are about 70% larger than those for the cumulative probit model. Even with such a huge sample size, the deviances cannot discriminate between the models and indicate that one fits and the other does not. The $\{\hat{\tau}_i\}$ for the row effects model are monotone decreasing for each link function but depart slightly from a linear trend as a function of the row numbers. Of the four pairs of adjacent categories for educational degree, the effects are a bit greater comparing less than high school and high school categories and comparing junior college and bachelor categories than comparing the other two pairs. In practical terms, though, the departure from a linear trend is not great. For simplicity, it is adequate to use the linear trend model even though it exhibits some lack of fit.

5.3 CUMULATIVE LOG-LOG LINKS: PROPORTIONAL HAZARDS MODELING

The *type I extreme value distribution*, sometimes called the *Gumbel distribution*, has cumulative distribution function

$$G(y) = \exp\left[-\exp\left(-\frac{y-a}{b}\right)\right],$$

where a is a location parameter and $b > 0$ is a scale parameter. Its mode is a (the mean is $a + 0.577b$) and the standard deviation is $1.283b$. The term *extreme value* refers to this being the limit distribution of the maximum of a sequence of independent and identically distributed continuous random variables. The distribution is often used to model extremes, such as the highest level of a river at a particular location over a year period.

The shape of the probability density function corresponding to this cdf is highly skewed to the right. Thus, the cdf approaches 1 at a much slower rate than it departs from 0, whereas the complement of the cdf approaches 1 at a much faster rate. The inverse of this cdf is the log-log link function.

5.3.1 Complementary Log-Log Link Function

An underlying extreme value distribution for a latent variable Y^* implies a cumulative link model for the observed ordinal response Y of the form

$$\log\{-\log[1 - P(Y \leq j)]\} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}. \tag{5.3}$$

This model applies the $\log(-\log)$ link function to the complement of the cumulative probabilities. The link function for the cumulative probability is called the *complementary log-log link*.

With this link function, $P(Y \leq j)$ approaches 1.0 at a faster rate than it approaches 0.0. This differs from models that use the probit or logit link function. For them, the link function $h(u)$ is symmetric: For any $0 < P(Y \leq j) < 1$, the link function applied to the cumulative probability satisfies

$$h[P(Y \leq j)] = -h[1 - P(Y \leq j)],$$

and $P(Y \leq j)$ approaches 1.0 at the same rate as it approaches 0.0. For the related *log-log link*, which is $\log\{-\log[P(Y \leq j)]\}$, $P(Y \leq j)$ approaches 1.0 at a slower rate than it approaches 0.0. It is appropriate when the complementary log-log link holds for the categories listed in the reverse order.

The model (5.3) with complementary log-log link function has the property

$$P(Y > j | \mathbf{x}_1) = [P(Y > j | \mathbf{x}_2)]^{\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]}.$$

For example, suppose that $\mathbf{x}_1$ is identical to $\mathbf{x}_2$ except that x_k is increased by 1. Then

$$P(Y > j \mid \mathbf{x} \text{ with } x_k = x + 1) = P(Y > j \mid \mathbf{x} \text{ with } x_k = x)^{\exp(\beta_k)}.$$

As x_k increases with fixed values for other predictors, $P(Y > j)$ increases or decreases according to whether β_k is negative or positive. The latent variable model actually implies that the model for Y has the form

$$\log\{-\log[1 - P(Y \leq j)]\} = \alpha_j - \boldsymbol{\beta}'\mathbf{x}.$$

The sign of β then reverses, so that *positive* β_k corresponds to *increasing* values of $P(Y > j)$, as in the usual sense of a positive association.

5.3.2 Example: Life Table for Gender and Race

Table 5.2 shows the life-length distribution for U.S. residents in 2004, by race and gender. Life length uses five ordered categories. The underlying continuous cdf of life length increases slowly at small to moderate ages but increases sharply at older ages. This suggests the complementary log-log link function. This link also results from assuming that the hazard rate increases exponentially with age, which happens for an extreme value distribution. The distributions shown are estimated population distributions based on census data. Sample sizes were unspecified and the samples were probably not simple random samples, so we will not use formal inference methods.

For gender g (1 = female; 0 = male), race r (1 = black; 0 = white), and life length Y, Table 5.2 also contains fitted distributions for the model

$$\log\{-\log[1 - P(Y \le j)]\} = \alpha_j + \beta_1 g + \beta_2 r.$$

We obtained these by fitting the model to the estimated population distributions shown in the table. The model describes well the four distributions, as indicated by the closeness of the fitted distributions shown in Table 5.2. One way to summarize the difference between two such distributions is with the *dissimilarity index*, which is half the sum of absolute differences between the fitted and estimated population distributions. This index takes values 0.004, 0.003, 0.0035, and 0.005 for the four groups. Another indication of the good fit is that if the model had been fitted to multinomial samples of size 1000 for each of the four groups that had the same percentages as Table 5.2 shows, the deviance would equal 2.1.

The fitted values correspond to model parameter values of $\beta_1 = -0.538$ and $\beta_2 = 0.611$. To interpret the gender effect, we note that the fitted cdf's satisfy

$$P(Y > j \mid G = 0, R = r) = [P(Y > j \mid G = 1, R = r)]^{\exp(0.538)}.$$

Given race, the proportion of men living longer than a fixed time equaled the proportion for women raised to the power of $\exp(0.538) = 1.71$. Given gender,

TABLE 5.2. Observed and Fitted (in Parentheses) Life-Length Distributions of U.S. Residents, as Percentages

| Gender | Race | Life Length | | | | |
		0–20	20–40	40–50	50–65	> 65
Female	Black	1.8 (1.6)	2.4 (2.7)	3.7 (3.5)	12.9 (13.1)	79.2 (79.1)
	White	0.9 (0.9)	1.3 (1.5)	1.9 (1.9)	8.0 (7.6)	87.9 (88.0)
Male	Black	2.6 (2.7)	4.9 (4.6)	5.6 (5.7)	20.1 (20.1)	66.8 (66.9)
	White	1.3 (1.5)	2.8 (2.5)	3.2 (3.3)	12.2 (12.3)	80.5 (80.4)

Source: 2008 Statistical Abstract of the United States, U.S. Census Bureau, Washington, DC.

the proportion of blacks living longer than a fixed time equaled the proportion of whites raised to the power of $\exp(0.611) = 1.84$. The β_1 and β_2 values (and the corresponding fitted distributions in Table 5.2) indicate that white men and black women had similar life-length distributions, that white women tended to have the longest lives, and that black men tended to have the shortest lives. If the probability of living longer than some fixed time equaled λ for white women, that probability was about $\lambda^{1.8}$ for white men and black women and about $\lambda^{3.5}$ for black men.

The cumulative logit model of proportional odds form also fits this life table well. (The deviance equals 2.4 when the model is fitted to counts having the estimated population distributions and a sample size of 1000 for each group.) Its gender effect is -0.604 and its race effect is 0.685. So if Ω denotes the odds of living longer than some fixed time for white women, the estimated odds of living longer than that time are $\exp(-0.604)\Omega = 0.55\Omega$ for white men, $\exp(-0.685)\Omega = 0.50\Omega$ for black women, and $\exp(-0.604 - 0.685)\Omega = 0.28\Omega$ for black men.

5.3.3 Proportional Hazards Model for Grouped Survival Times

In studies of survival, sometimes the survival-time response is measured by grouping the time scale into ordered categories. In Section 4.2.7 we noted that continuation-ratio logit models can describe hazards functions for grouped survival data. A certain model using the complementary log-log link function is also useful for such data. In fact, models using that link function are sometimes referred to as *proportional hazards models*, because the model results from generalizing the proportional hazards model for survival data to handle grouped survival times.

When a continuously measured survival time has probability density function f and cumulative distribution function F, the hazard function $h_t = f(t)/[1 - F(t)]$ is the instantaneous risk function. Incorporating explanatory variables, denote the hazard function by h_t when $\mathbf{x} = \mathbf{0}$ and by $h_t(\mathbf{x})$ otherwise. The proportional hazards model is

$$h_t(\mathbf{x}) = h_t \exp(\boldsymbol{\beta}'\mathbf{x}).$$

Equivalently, in terms of the survival functions $S_t = 1 - F(t)$ when $\mathbf{x} = \mathbf{0}$ and $S_t(\mathbf{x})$ otherwise,

$$S_t(\mathbf{x}) = S_t^{\exp(\boldsymbol{\beta}'\mathbf{x})}.$$

Now, for discretely measured survival, let $S_j = P(Y \geq j)$. As in Section 4.2, let

$$\omega_j = P(Y = j \mid Y \geq j) = 1 - \frac{S_{j+1}}{S_j}, \qquad j = 1, \ldots, c - 1.$$

Incorporating explanatory variables and letting ω_j and S_j without arguments refer to $\mathbf{x} = \mathbf{0}$, for the proportional hazards model,

$$\omega_j(\mathbf{x}) = 1 - \frac{S_{j+1}(\mathbf{x})}{S_j(\mathbf{x})} = 1 - \frac{S_{j+1}^{\exp(\boldsymbol{\beta}'\mathbf{x})}}{S_j^{\exp(\boldsymbol{\beta}'\mathbf{x})}} = 1 - (1 - \omega_j)^{\exp(\boldsymbol{\beta}'\mathbf{x})}.$$

It follows that

$$\log[-\log(1 - \omega_j(\mathbf{x}))] = \boldsymbol{\beta}'\mathbf{x} + \log[-\log(1 - \omega_j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}$$

with $\alpha_j = \log[-\log(1 - \omega_j)]$. That is, the model applies the complementary log-log link to the conditional probabilities used in continuation-ratio logits. For further details and related results, see Thompson (1977), Prentice and Gloeckler (1978), McCullagh (1980), and Aranda-Ordaz (1983). Perhaps surprisingly, this model for the conditional probabilities $\{\omega_j\}$ is equivalent to one using the same link function but with cumulative probabilities (Läärä and Matthews 1985).

5.3.4 Generalized Link Function Including Probit and Log-Log Links

Genter and Farewell (1985) introduced a generalized link function that permits comparison of fits provided by various cumulative link functions. Their generalized link function corresponds to the inverse cdf for a log-gamma density, which depends on a parameter q such that the density is positively skewed when $q < 0$, negatively skewed when $q > 0$, and is the standard normal density when $q = 0$. Special cases of the link function include the probit ($q = 0$), the complementary log-log ($q = 1$), and the log-log ($q = -1$).

The estimate $\hat{q}$ that maximizes the multinomial log-likelihood function provides an estimate of the best-fitting link function out of this generalized family. The test statistic for testing the adequacy of a particular link function equals double the difference between the maximized log-likelihood for the link function corresponding to the ML estimate $\hat{q}$ of q and the maximized log-likelihood for the chosen link function. Because the model with particular q is a special case of the model with unspecified q, this statistic has a null asymptotic chi-squared distribution with df $= 1$. As long as observations do not fall mainly in only one or two categories, Genter and Farewell (1985) found that Var($\hat{q}$) tends to decrease as the number of outcome categories c increases and as the effect size increases, thus making it easier to discriminate among the various link functions.

Suppose that we want to compare a particular pair of link functions, such as the probit and complementary log-log. The two models are not nested, so they cannot be compared with standard methods. However, if twice the difference between their maximized log likelihoods exceeds the appropriate percentile of the χ_1^2 distribution, we can conclude that the link with the smaller maximized likelihood fits more poorly. This is because the likelihood-ratio statistic comparing that link to the link corresponding to $\hat{q}$ would have an even larger test statistic. Thus, using the χ_1^2 distribution for this evaluation is conservative.

Unfortunately, the logit link is not a special case of this generalized link family. It is closely approximated by the probit link in this family. Lang (1999) proposed an alternative parametric family of link functions that includes the logit, log-log, and complementary log-log. He accomplished this by letting the cdf for the inverse link function be a mixture of the cdf's corresponding to these three link functions. He also considered a Bayesian analysis with this approach in which prior beliefs about

an appropriate link function could be combined with the data to obtain posterior information about an appropriate link function.

5.4 MODELING LOCATION AND DISPERSION EFFECTS

The cumulative link models studied so far in this chapter have the same effect for each cumulative probability. For this structure, settings of the explanatory variables are *stochastically ordered* on the response (recall Section 2.2.5): For any pair x_1 and x_2, either $P(Y \leq j \mid x_1) \leq P(Y \leq j \mid x_2)$ for all j or $P(Y \leq j \mid x_1) \geq P(Y \leq j \mid x_2)$ for all j. This is not surprising, because the latent variable construction showed that the model holds when an underlying continuous response has the usual regression model structure with a constant variance. In that case, the distribution of the response at different predictor values differs in terms of *location* but not *dispersion*.

5.4.1 Adding Dispersion Effects to the Cumulative Link Model

When a cumulative link model fits poorly, often it is because the dispersion changes considerably at different predictor values. For instance, perhaps responses tend to concentrate around a similar location for Y at x_1 as at x_2 but more dispersion occurs at x_1. In other words, at x_1 the responses concentrate more at the extreme categories than at x_2. Then it would not be surprising if $P(Y \leq j \mid x_1) > P(Y \leq j \mid x_2)$ for small j but $P(Y \leq j \mid x_1) < P(Y \leq j \mid x_2)$ for large j.

McCullagh (1980) generalized the cumulative link model to incorporate dispersion as well as location effects. With link function h, the model is

$$h[P(Y \leq j)] = \frac{\alpha_j - \beta' x}{\exp(\gamma' x)}. \tag{5.4}$$

The denominator contains scale parameters γ that describe how the dispersion depends on x. This model arises from a latent variable model in which the distribution of the latent variable has shape determined by h, such as normal for the probit link and logistic for the logit link. The latent variable has mean $\beta' x$ and standard deviation $\exp(\gamma' x)$ that varies as x does. (We use the negative coefficient for the $\beta' x$ term here to emphasize the connection between $\beta' x$ and the mean for the underlying latent variable model.)

The ordinary cumulative link model (5.1) is the special case of model (5.4) with $\gamma = 0$. Otherwise, at setting x, the cumulative probabilities tend to shrink toward their average when $\gamma' x > 0$. This creates higher probabilities in the end categories for Y and overall greater dispersion. The cumulative probabilities tend to move apart, creating less dispersion, when $\gamma' x < 0$.

5.4.2 Comparing Two Groups with Location and Dispersion Effects

Let's see how the cumulative link model for comparing two groups on an ordinal scale generalizes to permit dispersion effects. For this application, x consists of

a single binary predictor x represented by an indicator variable taking values 0 and 1. Model (5.4) with link function h simplifies to

$$h[P(Y \leq j)] = \alpha_j, \qquad x = 0,$$

$$h[P(Y \leq j)] = \frac{\alpha_j - \beta}{\exp(\gamma)}, \qquad x = 1.$$

The parameter β represents the difference between the means on the latent scale. The parameter $\exp(\gamma)$ represents the ratio of standard deviations for the two groups on the latent scale. When $\gamma > 0$, the group labeled $x = 1$ has more dispersion on Y than the group labeled $x = 0$.

To illustrate, consider the cumulative logit link. The case $\gamma = 0$ is the proportional odds form of the model, in which β is a location shift that determines a common cumulative log odds ratio for all 2×2 collapsings of a $2 \times c$ table. When $\gamma \neq 0$ the difference between the logits for the two groups, and hence the cumulative odds ratio, varies as j does.

Fitting model (5.4) is not straightforward, because it is not linear in the parameters. We can form the multinomial likelihood function by replacing response category probabilities by differences of cumulative probabilities, as in equation (3.13), then substituting $h^{-1}[(\alpha_j - \boldsymbol{\beta}'\mathbf{x})/\exp(\boldsymbol{\gamma}'\mathbf{x})]$ for the cumulative probabilities. The appendix to McCullagh (1980) derived likelihood equations and the information matrix. ML estimates can be obtained by using a nonlinear regression program to maximize the log likelihood, such as an iteratively reweighted Gauss–Newton algorithm (Cox 1995).

5.4.3　Example: Coronary Heart Disease and Smoking

In Section 3.6.5 we analyzed data on the relationship between the degree of coronary heart disease and smoking status. Table 5.3 shows the data again. The rows are stochastically ordered, but a cumulative logit model of proportional odds form fits poorly (deviance = 40.5, df = 3). In Section 3.6.5 we found that a nonproportional odds model for which the cumulative log odds ratio changes linearly across the response categories fits much better (deviance = 3.4, df = 2). That model estimated the cumulative log odds ratios as −1.02, −0.72, −0.42, and −0.12, close to the sample values of −1.04, −0.65, −0.46, and −0.07.

TABLE 5.3. Smoking Status and Degree of Coronary Heart Disease, with Percentages for Response in Parentheses

Smoking Status	Degree of Coronary Heart Disease[a]				
	0	1	2	3	4
Smoker	350 (22.6%)	307 (19.8%)	345 (22.3%)	481 (31.0%)	67 (4.3%)
Nonsmoker	334 (45.2%)	99 (13.4%)	117 (15.8%)	159 (21.5%)	30 (4.1%)

Source: Peterson and Harrell (1990), with permission.
[a]0, no disease; 4, very severe disease.

The varying cumulative log odds ratios also reflect differing dispersion for the two groups. The cumulative logit model having both location and dispersion effects with indicator $x = 1$ for smokers has $\hat{\beta} = 0.657$ (SE $= 0.077$) and $\hat{\gamma} = -0.308$(SE $= 0.054$). Compared to nonsmokers, responses for smokers tend to be located more toward the severe end of the response scale and show somewhat less dispersion. This model has deviance $= 6.8$ (df $= 2$), a decrease of 33.7 from the model having only a location effect, for which $\hat{\beta} = 0.737$.

5.4.4 Example: Vision Quality for Men and Women

Table 5.4 from Stuart (1953) shows data on assessment of right-eye vision for men and for women. From the percentages shown, men and women are not stochastically ordered. Relatively more men tend to fall at both the highest and lowest levels. Hence, a model having only a location parameter fits poorly. The cumulative logit model of proportional odds form with a binary indicator variable for gender has deviance $= 128.4$(df $= 2$). Its location parameter estimate of 0.038(SE $= 0.039$) seems to suggest a lack of difference between the groups.

McCullagh (1980) and Cox (1995) analyzed the data with location and scale models. The special case of the cumulative logit model with a dispersion effect but with location effect $\beta = 0$ fits well (deviance $= 2.6$, df $= 2$). With an indicator variable that equals 1 for females, the estimate $\hat{\gamma} = -0.271$ (SE $= 0.025$) reflects the smaller dispersion for the female responses. For an underlying continuous distribution of right-eye quality, the means seem to be similar for women and men, but the ratio of standard deviations is estimated to be $\exp(-0.271) = 0.76$.

5.5 ORDINAL ROC CURVE ESTIMATION

Diagnostic tests are used to detect many undesireable medical conditions, such as a disease of a particular type. For example, some diagnostic tests use x-rays or other imaging devices such as the mammogram (for diagnosing breast cancer) and the MRI body scan. A diagnostic test result is called *positive* if it states that the disease is present and *negative* if it states that the disease is absent. The accuracy of a diagnostic test is often assessed by two conditional probabilities:

$$\text{sensitivity} = P(\text{positive result} \mid \text{disease present}),$$

$$\text{specificity} = P(\text{negative result} \mid \text{disease absent}).$$

TABLE 5.4. Quality of Right-Eye Vision by Gender, with Conditional Distribution of Vision Quality in Parentheses

| Gender | Quality of Right-Eye Vision | | | |
	0 (Highest)	1	2	3 (Lowest)
Males	1053 (32.5%)	782 (24.1%)	893 (27.5%)	514 (15.9%)
Females	1976 (26.4%)	2256 (30.2%)	2456 (32.8%)	789 (10.6%)

Source: Stuart (1953), with permission of the Biometrika trustees.

The higher these two probabilities, the better the diagnostic test. A *false positive* occurs when the subject does not have the disease but the test is positive, which happens with rate

false positive rate $= P$(positive result | disease absent) $= 1 -$ specificity.

5.5.1 Ordinal Sensitivity and Specificity

Often, a diagnostic rating Y has an ordinal scale, such as $1 =$ definitely normal, $2 =$ probably normal, $3 =$ equivocal, $4 =$ probably abnormal, and $5 =$ definitely abnormal. Let x be the indicator of actual disease status, with $x = 1$ when the disease is present and $x = 0$ when the disease is absent. If we regard outcome $Y > j$ on the ordinal scale as being a positive response, then

specificity $= P(Y \leq j \mid x = 0)$,   sensitivity $= P(Y > j \mid x = 1)$.

A *receiver operating characteristic* (ROC) *curve* is a graphical way to summarize the performance of a diagnostic test. For various criteria for calling a diagnostic test result positive, the curve plots the false positive rate on the horizontal axis and the sensitivity on the vertical axis. For an ordinal response, the point on the ROC curve corresponding to the definition $Y > j$ for a positive outcome has coordinates

$[1 -$ specificity, sensitivity$] = [P(Y > j \mid x = 0), \ P(Y > j \mid x = 1)]$.

The ROC curve is constructed by plotting these points for $j = 0, 1, \ldots, c$. The curve connects the point $(0, 0)$, which occurs for $j = c$, with the point $(1, 1)$, which occurs when $j = 0$. For j between 1 and $c - 1$, the points usually all fall above the straight line connecting the points $(0, 0)$ and $(1, 1)$. If a point falls below that line, then for some definition of "positive," predictions are better by guessing randomly than by using the diagnostic test.

5.5.2 Cumulative Link Models and ROC Curves

Suppose that we use a cumulative link model

$$h[P(Y \leq j)] = \alpha_j - \beta x$$

to describe the impact of x on Y, using data for which we can measure both variables. We use the negative coefficient for the βx term here so that $\beta > 0$ corresponds to having a greater likelihood of a positive response when the disease is present than when it is absent.

Based on the model fit, the point for the ROC curve that is plotted when category j is the cutoff point for a positive outcome is

$$[\hat{P}(Y > j \mid x = 0), \ \hat{P}(Y > j \mid x = 1)] = [1 - h^{-1}(\alpha_j), \ 1 - h^{-1}(\alpha_j - \beta)].$$

For the logit link, for example, $\exp(\hat{\beta})$ is the estimated odds of a positive response for a diseased patient divided by the estimated odds of a positive response for a nondiseased patient. This is true for each possible cutoff point for a positive outcome. Based on the model fit, the point for the ROC curve that is plotted when category j is the cutoff point for a positive outcome is

$$[\hat{P}(Y > j \mid x = 0),\ \hat{P}(Y > j \mid x = 1)] = \left[\frac{1}{1 + \exp(\hat{\alpha}_j)},\ \frac{1}{1 + \exp(\hat{\alpha}_j - \hat{\beta})}\right].$$

For the probit link,

$$[\hat{P}(Y > j \mid x = 0),\ \hat{P}(Y > j \mid x = 1)] = [1 - \Phi(\hat{\alpha}_j),\ 1 - \Phi(\hat{\alpha}_j - \hat{\beta})].$$

In practice, rather than plotting only the $c - 1$ pairs of these estimated probabilities for $j = 1, \ldots, c - 1$, a smooth curve is constructed by letting $\hat{\alpha}_j$ vary continuously over the real line for the given $\hat{\beta}$. With a symmetric link function h such as the probit or logit, for which $h[P(Y \leq j)] = -h[1 - P(Y \leq j)]$, the ROC curve has a symmetric appearance. The curve approaches the point (1,1) with the same shape as it approaches the point (0,0), that is, symmetry about the line drawn from the top left to the bottom right of the graph. For the cumulative logit model of proportional odds form or the corresponding cumulative probit model, the ROC curve is necessarily concave when $\hat{\beta} > 0$.

Figure 5.1 shows the ROC curve for the cumulative probit model for four possible values of $\hat{\beta}$. For example, let $\hat{\alpha}_j = 0.0$, for which the false positive rate
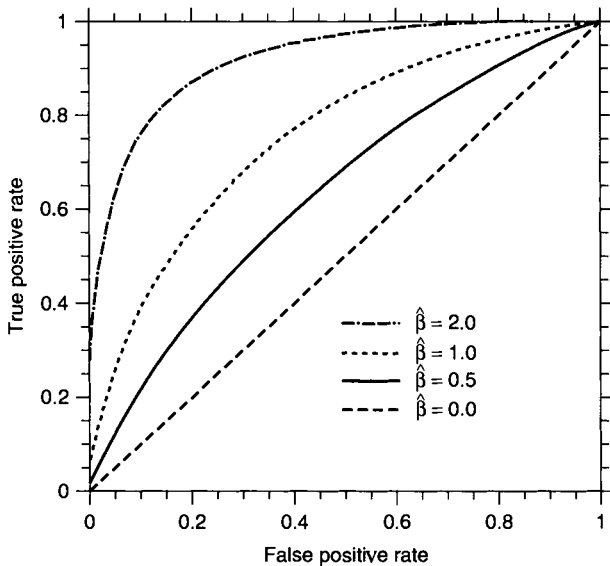
Figure 5.1. ROC curves for cumulative probit model with four different disease indicator effects $\hat{\beta}$. (From Tosteson and Begg (1988), with permission of Sage Publications.)

on the horizontal axis is $1 - \Phi(0) = 0.50$. Then the true positive rate on the vertical axis is $1 - \Phi(-\hat{\beta}) = \Phi(\hat{\beta})$, which for $\hat{\beta} = (0.0, 0.5, 1.0, 2.0)$ equals $(0.50, 0.69, 0.84, 0.98)$. As $\hat{\beta}$ increases, reflecting a better diagnostic process, the area under the ROC curve increases.

In addition to the disease indicator x, the model for the diagnostic rating Y can contain other explanatory variables that may have effects on the diagnosis, such as whether or not the patient has certain symptoms. Tosteson and Begg (1988) showed that the same ROC curves then result as when the model does not contain such variables, because the extra terms can be absorbed into the intercept term. Changing levels for an additional predictor corresponds to movement along the ROC curve rather than creation of a new curve. For example, if a binary predictor is an indicator for two different raters, the effect of its coefficient is merely to translate by that constant the cutpoints for one rater relative to the other one. However, when the additional predictors include an interaction between an explanatory variable and the disease status variable, a different ROC curve occurs at each setting of those variables.

5.5.3   Dispersion Effects and Area Under the ROC Curve

More generally, the ROC curve can be based on a model that also includes dispersion effects, such as model (5.4), which is

$$h[P(Y \leq j)] = \frac{\alpha_j - \boldsymbol{\beta}'\mathbf{x}}{\exp(\boldsymbol{\gamma}'\mathbf{x})}.$$

Tosteson and Begg (1988) suggested that this generalized model produces shapes for ROC curves that better resemble sample ROC plots often seen in practice, such as curves that are not symmetric or even concave. They noted that for such a model fitted to radiologic data, the estimated dispersion term $\exp(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i)$ for those with the disease often exceeds 1. That is, the spread of responses for diseased subjects (actual disease status $x = 1$) is greater than that for nondiseased subjects (disease status $x = 0$). This may reflect the fact that healthy physiology does not vary as much in its radiologic image as does abnormal physiology.

Consider the special case of the generalized model with the probit link function and with the disease status indicator as the only predictor in both portions of the model, with location effect β and dispersion effect γ. Then the area under the ROC curve is

$$\text{area} = \Phi\left(\frac{\beta}{\sqrt{1 + e^{2\gamma}}}\right).$$

See Tosteson et al. (1994), who also derived a standard error for its ML estimate. When $\gamma = 0$, the area is $\Phi(\beta/\sqrt{2})$. When $\beta = 0$ also, the ROC curve is the line with intercept 0 and slope 1, and the area under the curve is 0.50. For fixed γ, the area under the ROC curve is monotone increasing in β, with limiting value 1 as $\beta \to \infty$. For fixed β, the area under the ROC curve is monotone decreasing in γ, with limiting value 0.50 as $\gamma \to \infty$. Let Y_1 denote a random observation for a subject with the disease and Y_2 a random observation for a subject without

the disease. Then the area under the ROC curve equals the value of the stochastic superiority measure $\alpha = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ introduced in Section 2.1.4 (Bamber 1975).

Including a covariate term in the dispersion part of the model has the effect of either raising the ROC curve or lowering it. Including interaction terms with the disease status and explanatory variables provides the flexibility of different shapes for the ROC curve.

5.5.4 Example: Ultrasonography Cancer Detection

Table 5.5, based on an example discussed by Tosteson and Begg (1988), refers to the use of ultrasonography in detecting the presence or absence of hepatitic metasteses in patients with primary cancers of either the breast or the colon. Let x_1 be the indicator of hepatic metasteses (1 = yes, 0 = no) and let x_2 indicate the type of cancer (1 = breast, 0 = colon). The cumulative probit model

$$\Phi[P(Y \le j)] = \frac{\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \beta_3(x_1 x_2)}{\exp[\gamma_1 x_1 + \gamma_2 x_2 + \gamma_3(x_1 x_2)]}$$

has the estimates shown in Table 5.6. Figure 5.2 shows the ROC curves for the two types of cancer corresponding to these estimates. The curve for breast cancer

TABLE 5.5. Example of Ultrasound Rating Data for Breast Cancer and Colon Cancer

Hepatic Metasteses	Tumor Site	Ultrasonography Rating				
		1	2	3	4	5
No	Colon	47	17	2	0	0
Yes	Colon	4	1	2	2	13
No	Breast	6	5	2	1	0
Yes	Breast	0	2	0	2	5

Source: Based on an example described in Tosteson and Begg (1988).

TABLE 5.6. ML Estimates for Cumulative Probit Models Fitted to Ultrasonography Data

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
Original model						
Estimate	2.64	0.23	−0.74	1.81	0.42	−1.28
SE	0.85	0.24	1.02	0.54	0.38	0.69
Reduced model[a]						
Estimate	2.18	—	—	1.26		
SE	0.47	—	—	0.44		

[a]Reduced model has location and dispersion effect only for hepatic metasteses indicator.
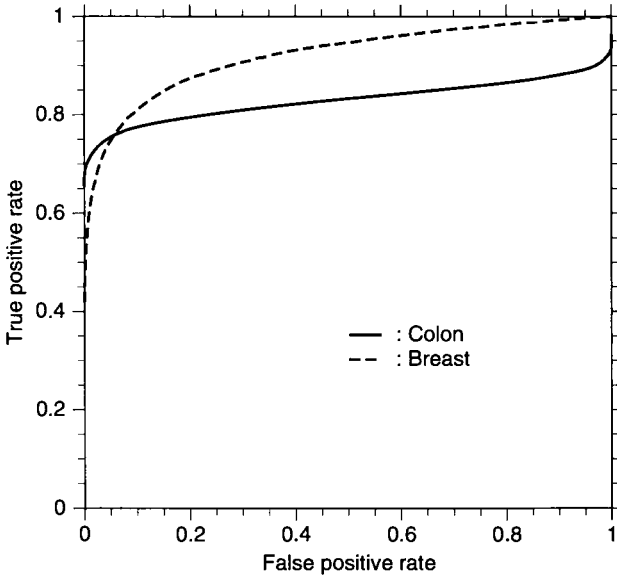
Figure 5.2. Estimated ROC curve for colon and breast cancer metasteses, based on cumulative probit model, including dispersion effect. (From Tosteson and Begg (1988), with permission of Sage Publications.)

suggests that when its false positive rate is high, that rate may even exceed the true positive rate.

The standard errors for the coefficients of x_2 and $x_1 x_2$ in the location and scale parts of the model suggest that it may be adequate to use a simpler model with x_1 alone in both parts, thus using the same curve for each type of cancer. This is verified by the likelihood-ratio statistic comparing the model fits, which equals 6.5 with df = 4. The simpler model has deviance = 11.55 (df = 10), $\hat{\beta}_1 = 2.18$ (SE = 0.47), and $\hat{\gamma} = 1.26$ (SE = 0.44). The positive $\hat{\gamma}$ estimate suggests that greater dispersion exists when hepatic metastases are present.

5.6   MEAN RESPONSE MODELS

In Chapters 3 to 5 we have introduced a variety of models for the outcome probabilities of an ordinal response. Cumulative link models apply link functions to cumulative probabilities. Adjacent-categories models and the stereotype model apply link functions to conditional probabilities, given occurrence in a pair of categories. Continuation-ratio models apply link functions to conditional probabilities, given occurrence above some category or given occurrence below some category. In this section we present a model that differs from models studied previously in that it describes a single summary of the outcome probabilities, the expected response, rather than the outcome probabilities themselves. The formula for the expected response resembles the ordinary regression formula for a quantitative response variable.

The model requires assigning monotone scores $v_1 \leq v_2 \leq \cdots \leq v_c$ to the outcome categories. At each fixed setting of explanatory variables $\mathbf{x}$, let

$$\mu(\mathbf{x}) = \sum_j v_j \pi_j(\mathbf{x})$$

denote the mean response on Y for those scores. The model

$$\mu(\mathbf{x}) = \alpha + \boldsymbol{\beta}'\mathbf{x} \tag{5.5}$$

assumes a linear relationship between the mean and the explanatory variables. We refer to this class of models as *mean response models*.

5.6.1 Fitting Mean Response Models

Assume that the observations on Y at different values of $\mathbf{x}_i$ are independent multinomial samples. Bhapkar (1968) and Grizzle et al. (1969) used weighted least squares (WLS) to fit mean response models. Since the outcome probabilities change as $\mathbf{x}$ changes, so does the variance of Y, and the WLS approach weights each sample mean by the inverse of its estimated variance. This requires all explanatory variables to be categorical, because many observations must occur at each predictor value in order to estimate the variance. In practice, this means that the overall sample size must be relatively large and the data cannot be sparse.

The ML approach for maximizing the product multinomial likelihood is more general than WLS. It applies for either categorical or continuous explanatory variables, and it does not require a large sample. Haber (1985), Lipsitz (1992), and Lang (2004, 2005) presented algorithms for ML fitting of families of models that include mean response models. The fitting process is somewhat complex, because the probabilities in the multinomial likelihood function are not direct functions of the parameters in model (5.5). The ML fit produces estimated response probabilities at each setting for the explanatory variables that maximize the product multinomial likelihood function under the constraint that they satisfy the model. Specialized software is available[1] that can fit a very broad class of models that includes mean response models.

5.6.2 Example: Political Ideology by Political Party and Gender

We illustrate the mean response model using Table 5.7, from the 2006 General Social Survey. We model the mean of $Y =$ political ideology using gender and political party identification. For simplicity, we use political ideology scores that are the category numbers. The sample means for the three party identifications (Democrat, Independent, Republican) are (3.54, 3.98, 4.96) for females and (3.52, 4.01, 5.14) for males. For each gender, responses tend to be more conservative for Republicans than for the other two party IDs.

[1] The mph.fit R function described in the Appendix, as illustrated at www.stat.ufl.edu/~aa/ordinal/ord.html.

TABLE 5.7. Political Party Identification and Political Views

	Party ID	1	2	3	4	5	6	7	Total
					Political Views[a]				
Females	Democrat	42	201	136	320	83	63	18	863
	Independent	33	87	107	459	123	92	19	920
	Republican	5	19	29	177	121	183	52	586
Males	Democrat	28	120	89	202	51	37	10	537
	Independent	20	79	124	362	120	90	18	813
	Republican	3	12	26	128	107	211	47	534

Source: 2006 General Social Survey.
[a] 1, extremely liberal; 2, liberal; 3, slightly liberal; 4, moderate; 5, slightly conservative; 6, conservative; 7, extremely conservative.

Let g be an indicator variable for gender (1 = females, 0 = males) and let p_1 and p_2 be indicator variables for political party identification ($p_1 = 1$ for Democrats and 0 otherwise, $p_2 = 1$ for Independents and 0 otherwise, $p_1 = p_2 = 0$ for Republicans). The mean response model with main effects but no interaction has ML fit,

$$\hat{\mu} = -5.081 - 0.063g - 1.513p_1 - 1.049p_2.$$

For a given political party, there seems to be essentially no difference in mean political ideology for males and females. For a given gender, the mean political ideology for Republicans is estimated to be about one category more conservative than for Independents and about 1.5 categories more conservative than for Democrats.

The goodness-of-fit statistics are $G^2 = 4.18$ and $X^2 = 4.17$. Since sample means occur at six party ID × gender combinations and the model has four parameters, the residual df = 2. The fit seems adequate (P-value = 0.12). The SE values are 0.040 for the g effect, 0.052 for the p_1 effect, and 0.048 for the p_2 effect. In the population, we can be 95% confident that the difference between the mean political ideology for Democrats and for Republicans falls in the interval $-1.513 \pm 1.96(0.052)$ for each gender. This rounds to $(-1.6, -1.4)$, quite close to 1.5 categories more liberal for Democrats.

5.6.3 Advantages and Disadvantages of Mean Response Models

Treating ordinal variables in a quantitative manner is sensible if their categorical nature reflects crude measurement of an inherently quantitative variable. In fact, we have seen that ordinal cumulative link models result from latent variable mean response models. Mean response models for ordinal categorical response variables provide yet another way of approximating ordinary regression models for latent response variables that we would ideally like to observe.

With $c = 2$, without loss of generality we can take $v_1 = 0$ and $v_2 = 1$. The model then specifies that the probability in a particular outcome category is a linear function of the predictor variables. For binary outcomes, that model is called the *linear probability model*. With multiple predictors, such a model is rarely adequate, because of the restricted [0, 1] range for probabilities. Often, ordinary ML fitting fails, because the iterative process generates an estimated probability outside the [0, 1] range for at least one predictor value.

With $c > 2$, ML fitting of the mean response model can also have difficulties, because the mean response must fall between v_1 and v_c. In addition, the example in Section 1.3.3 illustrated that having upper and lower bounds for the observed response can cause floor effects and ceiling effects that bias the results. This tends to be less problematic as c increases and there is reasonable dispersion of responses over the c categories throughout the domain of interest for the explanatory variables. Fitting is most likely to encounter problems when a relatively high proportion of observations falls in category 1 or in category c of Y.

With $c > 2$, the mean response model does not specify the response probabilities structurally but merely describes the dependence of the mean on x. That is, unlike models considered previously, specifying parameters for a mean response model does not uniquely determine cell probabilities. Thus, mean response models do not directly specify structural aspects, such as stochastic orderings. These models do not represent the categorical response structure as fully as do models for probabilities, and conditions such as independence do not occur as special cases.

Although mean response models have these severe limitations, they have the advantage of providing simple descriptions. Effects are described by slopes or differences between means instead of by odds ratios or parameters in cumulative link models. As c increases, mean response models also interface with ordinary regression models for quantitative response variables. For moderate to large c, the mean response model approximates results for the regression model that would be appropriate if we could measure Y in a truly quantitative manner, with ungrouped data.

CHAPTER NOTES

Section 5.1: Cumulative Link Models

5.1. Other articles that discussed cumulative link models include Cowles (1996), Ishwaran and Gatsonis (2000), and Chen and Dey (2000) using MCMC methods for Bayesian model fitting. For latent variable modeling with a set of ordinal response variables for various link functions, see Bartholomew (1983) for an early review and Moustaki (2000) for later work. An alternative latent variable approach to deriving an ordinal model is based on maximizing random utility. Small (1987) presented this approach in the context of discrete choice modeling. Yee and Wild (1996) defined generalized additive models for ordinal responses. Such models are especially useful for smoothing ordinal response data having continuous explanatory variables without assuming linearity of effects. In Note 11.2 we refer to other ways of smoothing ordinal data.

Section 5.2: Cumulative Probit Models

5.2 Early uses of the cumulative probit model were by Aitchison and Silvey (1957), Ashford (1959), Gurland et al. (1960), Bock and Jones (1968, Chap. 8.1), and Samejima (1969). Bock (1975, Sec. 8.1.6) and McKelvey and Zavoina (1975) motivated the model by the regression model for an underlying normal latent variable, extending models of Aitchison and Silvey (1957) and Ashford (1959) for a single predictor. Later uses include Muthén (1984) for latent variable models, Tosteson et al. (1989) for a measurement error model, Agresti (1992a) for paired preference data, Becker and Kennedy (1992) for a graphical exposition, Hausman et al. (1992) for modeling transaction stock prices, Weiss (1993) and Kim (1995) for modeling a bivariate response, Saei et al. (1996) for modeling repeated measures of count data, Cowles (1996) using MCMC methods, Ronning and Kukuk (1996) for multivariate modeling assuming an underlying joint normal distribution, and Glewwe (1997) for a test of the normality assumption for the latent variable model. The notes for Chapters 10 and 11 list several references that deal with multivariate cumulative probit models.

Section 5.3: Cumulative Log-Log Links: Proportional Hazards Modeling

5.3 Farewell (1982) generalized the complementary log-log model to allow variation among the sample in the category boundaries for the underlying scale by letting $\exp(\alpha_1)$ vary among the sample according to a gamma distribution, with $\alpha_j - \alpha_1$ the same for all subjects. This type of model relates to random effects models having random intercepts (Section 1.3) for which the variance component describes subject heterogeneity. Other articles in which proportional hazards models were discussed include Läärä and Matthews (1985), Nandram (1989), Barnhart and Sampson (1994), Crouchley (1995), Cowles (1996), Ten Have (1996), Hedeker et al. (2000), and Grilli (2005).

Section 5.4: Modeling Location and Dispersion Effects

5.4 Other articles in which modeling dispersion as well as location were considered include Nair (1987), Tosteson and Begg (1988), Tutz (1989), Hamada and Wu (1990), and Williams (2009). Cox (1995) presented a general model that contains as special cases the cumulative logit model having dispersion effects and the cumulative logit model (3.15) allowing partial proportional odds. Cox proposed ML estimation with a nonlinear regression program using the Gauss–Newton method (such as PROC NLIN in SAS), employing constraints so that estimated cumulative probabilities are not out of order. He presented several examples, including an alternative analysis of Table 5.3.

Section 5.5: Ordinal ROC Curve Estimation

5.5 For more details about using cumulative link models with location and dispersion terms to construct ROC curves, see Tosteson and Begg (1988) and Tosteson

et al. (1994). Lui et al. (2004) proposed methods for testing equality between two diagnostic procedures with paired-sample ordinal data that can be stratified by cases and noncases. One method is based on correctly identifying the case for a randomly selected pair of a case and a noncase, and the other is based directly on the sensitivity and specificity. Toledano and Gatsonis (1996) and Ishwaran and Gatsonis (2000) extended ordinal ROC curve analysis for data in which several raters analyze the same cases. Waegemana et al. (2008) discussed ROC analysis in a machine learning context.

Section 5.6: Mean Response Models

5.6 Articles that discussed mean response models for ordered categorical response variables include Yates (1948), Bhapkar (1968), Grizzle et al. (1969), Koch and Reinfurt (1971), Williams and Grizzle (1972), Koch et al. (1977), Meeks and D'Agostino (1983), Haber (1985), Agresti (1986) for an R^2 measure, Agresti (1992), Lang et al. (1999) for comparing mean responses of multivariate ordinal data, and Haber (1985), Lipsitz (1992), and Lang (2004, 2005) for ML fitting.

EXERCISES

5.1. For the cumulative probit model $\Phi^{-1}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}$, explain why a 1-unit increase in x_k corresponds to a β_k standard deviation increase in the expected underlying latent response, controlling for other predictors.

5.2. Refer to the example in Section 1.3.3. Generate 100 observations from the given latent variable model, which has $E(y^*) = 20 + 0.60x - 40.0z$.

   (a) Plot the data and fit the model using OLS.
   (b) Categorize y^* into y using five categories, as in that example. Now using y as the response, fit the same model as well as the extended model allowing interaction, using OLS with the scores (1, 2, 3, 4, 5). In the interaction model, compare the slopes for the two levels of z.
   (c) Fit the ordinal probit model to y. Show how results compare to those for the model for y^*, and show that there is actually no need for an interaction term.

5.3. For cumulative link model (5.1), show that for $1 \leq j < k \leq c - 1$, $P(Y \leq k \mid \mathbf{x}) = P(Y \leq j \mid \mathbf{x}^*)$, where $\mathbf{x}^*$ is obtained by increasing the ith component of $\mathbf{x}$ by $(\alpha_k - \alpha_j)/\beta_i$. Interpret.

5.4. A cumulative link row effects model for an $r \times c$ contingency table with a qualitative predictor is

$$h[P(Y \leq j)] = \alpha_j + \tau_i, \qquad j = 1, \ldots, c - 1.$$

   (a) Show that the residual df $= (r-1)(c-2)$.

   (b) When this model holds, explain why independence corresponds to $\tau_1 = \cdots = \tau_r$ and the test of independence has df $= r-1$.

   (c) When this model holds, explain why the rows are stochastically ordered on Y.

5.5. Let $F_1(y) = 1 - \exp(-\lambda y)$ for $y > 0$ be a negative exponential cdf with parameter λ, and let $F_2(y) = 1 - \exp(-\mu y)$ for $y > 0$. Show the difference between the cdf's on a complementary log-log scale is identical for all y. Give implications for data analysis with an ordered categorical response variable.

5.6. For a link function h, consider the model of form presented in Section 4.2,

$$h[\omega_j(\mathbf{x})] = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}, \qquad \text{where } \omega_j = \frac{\pi_j}{\pi_j + \cdots + \pi_c}.$$

   (a) Explain why the fit of this model is the same when fitted separately for $j = 1, \ldots, c-1$ or when fitted simultaneously.

   (b) For the complementary log-log link, show that this model is equivalent to one using the same link function for cumulative probabilities (Läärä and Matthews 1985).

5.7. Using the logit link, fit the model (5.4) with location and dispersion terms to Table 3.8 from Section 3.5.6 in Chapter 3, for which the cumulative logit model of proportional odds form has lack of fit. Interpret.

5.8. Table 5.8 summarizes observations of passengers in autos and light trucks involved in accidents in Maine in one year. The table classifies passengers

TABLE 5.8. Passenger Observations

Gender	Location	Seat Belt Use	Extent of Injury[a]				
			1	2	3	4	5
Female	Urban	No	7,287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3,246	73	710	159	31
		Yes	6,134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6,123	141	710	188	45
		Yes	6,693	74	353	74	12

Source: Cristanna Cook, Medical Care Development, Augusta, Maine.
[a] 1, not injured; 2, injured but not transported by emergency medical services; 3, injured and transported by emergency medical services but not hospitalized; 4, injured and hospitalized but did not die; 5, injured and died.

by gender, location of accident, seat belt use, and extent of injury. Find a cumulative link model that describes these data well. Describe the effects using estimated parameters from the model.

5.9. Refer to Exercise 2.7. Analyze these data using methods of this chapter.

CHAPTER 6

Modeling Ordinal Association Structure

Our focus so far in this book has been on modeling a single ordinal response variable. The remainder of the book deals with bivariate and multivariate ordinal responses. We focus next on the analysis of association between response variables, in this chapter using models and in Chapter 7 using non-model-based summary measures. When each response variable has the same categories, such as in longitudinal studies that measure a variable repeatedly over time, it is often of interest to compare and model the marginal distributions. In Chapter 8 we present ways of doing this and in Chapters 9 and 10 extend the analyses to more general models (such as including random effects), with emphasis on effects of explanatory variables.

For joint distributions of categorical response variables, *loglinear models* describe the dependence structure. For example, loglinear modeling can analyze whether the association between a pair of variables is homogeneous across the categories of other variables, and if so, whether those variables are conditionally independent. In this chapter we assume familiarity with standard loglinear models for contingency tables. We introduce specialized loglinear models for ordinal response variables as well as other models not having loglinear structure that can describe ordinal association, such as models for global odds ratios. We refer to the models of this chapter as *association models*.

6.1 ORDINARY LOGLINEAR MODELING

Most of this chapter deals with modeling two-way contingency tables. Denote the observed cell counts by $\{n_{ij}\}$ and the expected cell counts by $\{\mu_{ij}\}$. We assume a multinomial sample over the cells, with fixed sample size n and multinomial probabilities $\pi_{ij} = \mu_{ij}/(\sum_a \sum_b \mu_{ab}) = \mu_{ij}/n$. The ML estimates of loglinear

association and interaction parameters are identical to the ML estimates under the assumption that the cell counts are independent Poisson variates with expected values $\{\mu_{ij}\}$. Similarly, for inferential analyses about model effect parameters, equivalent results occur under the Poisson and multinomial sampling assumptions.

6.1.1 Loglinear Models of Independence and of Association

The loglinear model of statistical independence for an $r \times c$ table is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y.$$

For identifiability, the row and column terms satisfy constraints such as $\lambda_r^X = \lambda_c^Y = 0$. This model has residual df $= (r - 1)(c - 1)$. The general loglinear model that permits association is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

There are $(r - 1)(c - 1)$ linearly independent λ_{ij}^{XY} terms. For example, with the constraints that these association parameters equal 0 in the last row and in the last column, these parameters are log odds ratios for the 2×2 rectangular patterns of cells that use the cell in the last row and last column as a baseline,

$$\lambda_{ij}^{XY} = \log \frac{\mu_{ij}\mu_{rc}}{\mu_{ic}\mu_{rj}}.$$

This model is *saturated*, having as many parameters as cell count observations. Because of this, its residual df $= 0$ and $\{\hat{\mu}_{ij} = n_{ij}\}$.

This loglinear model and ordinary loglinear models for multiway contingency tables have a serious limitation—they treat all classifications as nominal. If the order of a variable's categories changes in any way, the fit is the same. For ordinal classifications, these models ignore the ordinality. The following example illustrates this point.

6.1.2 Example: Astrology Belief and Educational Attainment

In Section 3.2.5 we analyzed a 5×3 cross-classification from the 2006 General Social Survey on highest educational degree and opinion about astrology. Table 6.1 shows the data again. Consider the loglinear model of independence. Table 6.1 also contains its ML fitted values $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$. The statistics for testing goodness of fit are the likelihood-ratio (deviance) statistic, $G^2 = 107.6$, and the Pearson statistic, $X^2 = 103.5$, where

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \quad \text{and} \quad G^2 = 2 \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}. \quad (6.1)$$

TABLE 6.1. Education and Belief About Astrology with Fit of Independence Model and Its Standardized Residuals

Highest Degree	Astrology Is Scientific		
	Not At All	Sort of	Very
< High school	98 (138.3, −6.4)	84 (56.5, 4.6)	23 (10.2, 4.4)
High school	574 (614.1, −4.0)	286 (250.7, 3.7)	50 (45.2, 1.1)
Junior college	122 (114.7, 1.3)	44 (46.8, −0.5)	4 (8.4, −1.6)
Bachelor	268 (226.7, 5.3)	57 (92.6, −4.8)	11 (16.7, −1.6)
Graduate	148 (116.1, 5.5)	23 (47.4, −4.4)	1 (8.5, −2.8)

Source: 2006 General Social Survey.

Under the null hypothesis of independence, each statistic has an approximate chi-squared distribution with df = 8. The model fits poorly. Yet adding the ordinary association term makes the model saturated.

Table 6.1 also contains standardized residuals for the independence model fit. In a particular cell, the standardized residual equals $n_{ij} - \hat{\mu}_{ij}$ divided by the estimated standard error of that difference, assuming the independence model. The standardized residuals in the corners stand out. Their pattern indicates lack of fit in the form of a negative trend. Subjects who are more highly educated are less likely to think that astrology has some scientific basis.

Models for ordinal variables use association terms that permit trends of this type. The models are more complex than the independence model, yet unsaturated. Moreover, the models have parameters that lead to simple descriptions of the association and that provide improved power in statistical inference for detecting effects.

6.2 LOGLINEAR MODEL OF LINEAR-BY-LINEAR ASSOCIATION

We first consider the case in which both variables in a two-way contingency table are ordinal. A simple model with just one more parameter than the independence model can describe a positive or a negative trend association between those variables.

6.2.1 Bilinear Association Term and Ordinal Trends

Let $u_1 \leq u_2 \leq \cdots \leq u_r$ be ordered row scores and $v_1 \leq v_2 \leq \cdots \leq v_c$ be ordered column scores. The model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j, \tag{6.2}$$

with constraints such as $\lambda_r^X = \lambda_c^Y = 0$. This model is the special case of the general loglinear model in which the association term has the structured form, $\lambda_{ij}^{XY} = \beta u_i v_j$. It uses only one parameter to describe association, whereas the general

model uses $(r - 1)(c - 1)$ parameters. Independence between the two variables is the special case in which $\beta = 0$.

The term $\beta u_i v_j$ represents the deviation of $\log \mu_{ij}$ from the independence pattern. The deviation is linear in the Y scores $\{v_j\}$ at a fixed level of X and linear in the X scores $\{u_i\}$ at a fixed level of Y. In column j, for instance, the deviation is a linear function of X, having the form slope $\times$ score for X, with slope βv_j. Because of this property, model (6.2) is called the *linear-by-linear association model*. We abbreviate it by L $\times$ L. The model implies that the greatest departures from independence are in the four corners of the table. Haberman (1974) introduced this and more general models by decomposing the association term λ_{ij}^{XY} in a loglinear model into orthogonal components. Birch (1965) had suggested the linear-by-linear form. Goodman (1979a) was highly influential in investigating this and more general models introduced in Section 6.3, with scores that are themselves parameters.

For the 2×2 table using the cells intersecting rows a and c with columns b and d, direct substitution shows that the model satisfies

$$\log \frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \beta(u_c - u_a)(v_d - v_b). \qquad (6.3)$$

This log odds ratio is stronger as $|\beta|$ increases and for pairs of categories that are farther apart. The direction and strength of the association depend on β. When $\beta > 0$, Y tends to increase as X increases. Expected frequencies are larger than expected (under independence) in cells where X and Y are both high or both low. When $\beta < 0$, Y tends to decrease as X increases. When the data display a positive or negative trend, the L $\times$ L model fits better than the independence model.

6.2.2   Choice of Scores and the Local Odds Ratios

The choice of row scores and column scores in the model affects the interpretation of β. Simple interpretations result when $u_2 - u_1 = \cdots = u_r - u_{r-1}$ and $v_2 - v_1 = \cdots = v_c - v_{c-1}$. In that case the *local odds ratios* (2.4) for adjacent rows and adjacent columns all take the same value. Duncan (1979) and Goodman (1979a) called this case *uniform association*. Figure 6.1 portrays local odds ratios having uniform value. For unit-spaced scores such as $\{u_i = i\}$ and $\{v_j = j\}$, the common local odds ratio equals e^β.

For quantitative variables for which possible values have been grouped into ordered categories, it is often sensible to choose scores that approximate distances between midpoints of categories for the quantitative scale. For example, if alcohol consumption is measured in terms of number of drinks per day using the categories (0, less than 1, 1–2, 3–5, 6 or more), scores such as (0, 0.5, 1.5, 4, 7) are sensible. Then β represents the log odds ratios for unit distances in the X and Y directions. However, you do not need to regard scores as approximations for distances between categories or as reasonable scalings of ordinal variables for the model to be valid. The scores merely imply a certain pattern for the local odds ratios. If the L $\times$ L model fits well with equally spaced row and column scores, the uniform local odds
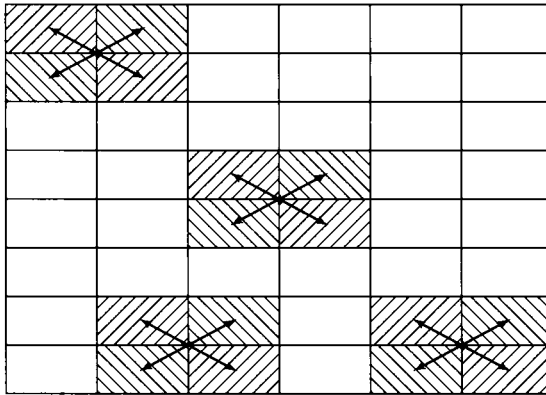
Figure 6.1. Uniform local odds ratios.

ratio describes the association regardless of whether the scores are sensible indexes of true distances between categories.

Two sets of scores having the same spacings, such as $\{u_1 = 1, u_2 = 2, u_3 = 3\}$ and $\{u_1 = -1, u_2 = 0, u_3 = 1\}$, yield the same ML estimate $\hat{\beta}$, the same fitted values $\{\hat{\mu}_{ij}\}$, and the same G^2 and X^2 values for testing fit. Two sets of scores with the same *relative* spacings (i.e., one set being a linear transformation of the other) yield the same fitted values but an appropriately rescaled $\hat{\beta}$. For instance, using row scores $\{u_i = 2i\}$ with $\{v_j = j\}$ yields the same fit as scores $\{u_i = i\}$ and $\{v_j = j\}$, but $\hat{\beta}$ and its SE are both half as large. Any linear transformation of the scores has no impact on inferential results or on model goodness of fit.

It is sometimes useful to standardize the scores, subtracting the mean and dividing by the standard deviation of the scores for the table marginal distributions. Then the standardized scores satisfy

$$\sum u_i \pi_{i+} = \sum v_j \pi_{+j} = 0,$$

$$\sum u_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1,$$

and similarly with sample marginal distributions. Then β represents the log odds ratio for a standard deviation increase in each variable. If the marginal standard deviations for the original scores are σ_X and σ_Y, and if β is the effect for the original scores and β^* is the effect for the standardized scores,

$$\beta^* = \beta \sigma_X \sigma_Y.$$

Often, it is sensible to regard the observed variables as crude measurement of underlying inherently continuous latent variables. The uniform association version of the $L \times L$ model tends to fit well when the underlying joint distribution is approximately bivariate normal and the marginal cutpoints for forming the categories are equally spaced. For standardized scores, β^* then approximates $\rho^*/[1 - (\rho^*)^2]$,

where ρ^* is the correlation between the underlying continuous variables. For weak associations, $\beta^* \approx \rho^*$. See Goodman (1981a,b, 1985), Wang (1987, 1997), and Becker (1989b) for further elaboration of this connection.

6.2.3 Corresponding Adjacent-Categories Logit Model

A logit formulation of the L × L model treats Y as a response variable and X as explanatory. Let $\pi_{j|i} = P(Y = j \mid X = i)$. Using logits for adjacent response categories (see Section 4.1) with model (6.2),

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \log \frac{\mu_{i,j+1}}{\mu_{ij}} = (\lambda^Y_{j+1} - \lambda^Y_j) + \beta(v_{j+1} - v_j)u_i.$$

For unit-spaced $\{v_j\}$, this simplifies to

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \alpha_j + \beta u_i,$$

where $\alpha_j = \lambda^Y_{j+1} - \lambda^Y_j$. The same linear logit effect β applies simultaneously for all $c - 1$ pairs of adjacent response categories: The odds that $Y = j + 1$ instead of $Y = j$ multiply by e^β for each unit change in x. In using equal-interval response scores $\{v_j\}$, we implicitly assume that the effect of x is the same on each of the $c - 1$ adjacent-categories logits for Y. This is the proportional odds structure for this logit model.

6.2.4 L × L Model Fitting and Inference

For independent Poisson sampling over the cells of a two-way contingency table, the kernel of the log-likelihood function is

$$L(\mu) = \sum_i \sum_j n_{ij} \log \mu_{ij} - \sum_i \sum_j \mu_{ij}.$$

This simplifies for the L × L model (6.2) to

$$L(\mu) = n\lambda + \sum_i n_{i+}\lambda^X_i + \sum_j n_{+j}\lambda^Y_j + \beta \sum_i \sum_j u_i v_j n_{ij}$$
$$- \sum_i \sum_j \exp(\lambda + \lambda^X_i + \lambda^Y_j + \beta u_i v_j).$$

Differentiating L separately with respect to $(\lambda^X_i, \lambda^Y_j, \beta)$ and setting the three partial derivatives equal to zero yields the likelihood equations

$$\hat{\mu}_{i+} = n_{i+}, \ i = 1, \ldots, r, \qquad \hat{\mu}_{+j} = n_{+j}, \ j = 1, \ldots, c,$$

$$\sum_i \sum_j u_i v_j \hat{\mu}_{ij} = \sum_i \sum_j u_i v_j n_{ij}.$$

Let $p_{ij} = n_{ij}/n$ and $\hat{\pi}_{ij} = \hat{\mu}_{ij}/n$. From the first two likelihood equations, $\hat{\pi}_{i+} = p_{i+}$ and $\hat{\pi}_{+j} = p_{+j}$ for all i and j. That is, the marginal distributions and hence the marginal means and variances for the chosen scores are identical for the fitted and observed distributions. The third likelihood equation is

$$\sum_i \sum_j u_i v_j \hat{\pi}_{ij} = \sum_i \sum_j u_i v_j p_{ij}.$$

Since the marginal distributions are identical for fitted and observed distributions, this equation implies that the correlation $\hat{\rho}$ between the scores for X and Y is also the same for both distributions. Let $\bar{u} = \sum_i u_i p_{i+}$ and $\bar{v} = \sum_j v_j p_{+j}$ denote the marginal sample means. The sample correlation $\hat{\rho}$ equals the covariance divided by the product of the sample standard deviations,

$$\hat{\rho} = \frac{\sum_i \sum_j (u_i - \bar{u})(v_j - \bar{v}) p_{ij}}{\sqrt{\left[\sum_i (u_i - \bar{u})^2 p_{i+}\right]\left[\sum_j (v_j - \bar{v})^2 p_{+j}\right]}}. \tag{6.4}$$

It falls between -1 and $+1$. The larger $\hat{\rho}$ is in absolute value, the farther the data fall from independence in a linear dimension. The fitted counts display the same positive or negative trend as the sample data, substituting $\hat{\pi}_{ij}$ for p_{ij} in (6.4).

Log-likelihood functions for loglinear models are concave, and the observed information matrix is identical to the expected information matrix. The Newton–Raphson iterative method, which for loglinear models is equivalent to Fisher scoring, yields the ML fit. This is easily implemented with software for generalized linear models, as the Appendix shows.

An infinite estimate occurs for $\hat{\beta}$ whenever either no pairs of observations are concordant or no pairs are discordant. Then the sufficient statistic $\sum_i \sum_j u_i v_j n_{ij}$ for β takes its minimum or maximum possible value given the row and column marginal counts. This happens when all observations fall in one row and in one column, with each at the highest or lowest level of a variable. It also happens when all observations fall on a diagonal of the table.

Since $\{u_i\}$ and $\{v_j\}$ are fixed, the L × L model (6.2) has only one more parameter (β) than the independence model. Its residual

$$\text{df} = rc - [1 + (r-1) + (c-1) + 1] = (r-1)(c-1) - 1 = rc - r - c.$$

The model is unsaturated for all but 2 × 2 tables. For large samples the Pearson X^2 and deviance G^2 statistics for testing the model fit by comparing $\{n_{ij}\}$ to the ML fitted counts $\{\hat{\mu}_{ij}\}$ have approximate chi-squared null distributions with $\text{df} = rc - r - c$.

Inference about the association parameter β uses standard methods. Wald or likelihood-ratio confidence intervals for β imply confidence intervals for odds ratios

such as local odds ratios. Significance tests of H_0: $\beta = 0$ are tests of independence that take into account the ordinality of the variables. Let $G^2(\text{I})$ denote the deviance statistic for testing the fit of the independence model, and let $G^2(\text{L} \times \text{L})$ denote the deviance for testing the fit of the linear-by-linear association model. The likelihood-ratio test statistic for H_0: $\beta = 0$ equals

$$G^2(\text{I} \mid \text{L} \times \text{L}) = G^2(\text{I}) - G^2(\text{L} \times \text{L}).$$

Designed to detect positive or negative trends, it has an asymptotic null chi-squared distribution with df $= 1$. Alternative chi-squared test statistics are the Wald statistic, $(\hat{\beta}/\text{SE})^2$, and the score statistic, which equals $(n - 1)\hat{\rho}^2$ for the sample correlation $\hat{\rho}$ between the scores for X and Y. (Section 7.3.4 presents the score test in a more general context.) For highly sparse data, another possibility is a bootstrap test using $G^2(\text{I} \mid \text{L} \times \text{L})$ for repeated samples from a multinomial with probabilities based on the fit of the independence model (Pettersson 2002).

When the L × L model holds, the ordinal test using $G^2(\text{I} \mid \text{L} \times \text{L})$ is asymptotically more powerful than the test using $G^2(\text{I})$. This is because for a fixed degree of effect (technically, the *noncentrality* for the noncentral chi-squared asymptotic distribution), the power of a chi-squared test increases when df decreases. When the L × L model holds, the noncentrality is the same for $G^2(\text{I} \mid \text{L} \times \text{L})$ and $G^2(\text{I})$; thus, $G^2(\text{I} \mid \text{L} \times \text{L})$ is more powerful, since it has df $= 1$ compared to df $= (r - 1)(c - 1)$ for $G^2(\text{I})$. The power advantage increases as r and c increase, because the noncentrality remains focused on df $= 1$ for $G^2(\text{I} \mid \text{L} \times \text{L})$ but df also increases for $G^2(\text{I})$.

6.2.5 Example: Astrology and Education Revisited

Now we continue analyzing Table 6.1 on highest educational degree and opinion about astrology. Table 6.2 shows the data again together with the fitted values for the linear-by-linear association model, using equally spaced scores for rows and columns. Table 6.3 shows software output. To obtain this, we added a quantitative variable to the linear predictor for the independence model having values equal to the product of row and column number (see the Appendix). Compared to the independence model, for which $G^2(\text{I}) = 107.6$ with df $= 8$, the L × L model fits dramatically better, with $G^2(\text{L} \times \text{L}) = 6.8$ based on df $= 7$. Comparing to Table 6.1, the improved fit is especially noticeable in the corners of the table, where the model predicts the greatest departures from independence.

With unit-spaced scores, $\hat{\beta} = -0.390$ (SE $= 0.042$). Subjects having higher levels of education tend to see less scientific basis to astrology. Then $\exp(\hat{\beta}) = \exp(-0.390) = 0.68$ is the estimated common local odds ratio. The strength of association seems weak. From (6.3), however, nonlocal odds ratios are stronger. The estimated odds ratio for the four corner cells equals

$$\exp[\hat{\beta}(u_5 - u_1)(v_3 - v_1)] = \exp[-0.390(5 - 1)(3 - 1)] = 0.044,$$

TABLE 6.2. Education and Belief About Astrology with Fit of Linear-by-Linear Association Model

Highest Degree	Astrology Is Scientific		
	Not At All	Sort of	Very
< High school	98 (104.6)	84 (78.4)	23 (22.0)
High school	574 (567.3)	286 (287.9)	50 (54.8)
Junior college	122 (122.5)	44 (42.1)	4 (5.4)
Bachelor	268 (268.2)	57 (62.4)	11 (5.4)
Graduate	148 (147.4)	23 (23.2)	1 (1.4)

TABLE 6.3. Output for Fitting Linear-by-Linear Association Model to Table 6.1

```
             Criteria For Assessing Goodness Of Fit
        Criterion              DF        Value        Value/DF
        Deviance               7        6.8389         0.9770
        Pearson Chi-Square     7        8.0767         1.1538

                         Standard     Wald 95%          Chi-
   Parameter     DF  Estimate    Error    Confidence Limits   Square
   Intercept      1    0.3153   0.4948   -0.6544    1.2850     0.41
   astro      1   1   -1.5585   0.1453   -1.8433   -1.2737   115.02
   astro      2   1   -0.2885   0.0821   -0.4495   -0.1276    12.35
   astro      3   0    0.0000   0.0000    0.0000    0.0000      .
   degree     0   1    4.3349   0.4672    3.4192    5.2506    86.08
   degree     1   1    4.8562   0.3656    4.1397    5.5728   176.44
   degree     2   1    2.1538   0.2672    1.6301    2.6775    64.98
   degree     3   1    1.7679   0.1559    1.4623    2.0735   128.58
   degree     4   0    0.0000   0.0000    0.0000    0.0000      .
   uv             1    0.3898   0.0421    0.3073    0.4723    85.75

                  LR Statistics For Type 3 Analysis
              Source          DF    Chi-Square   Pr > ChiSq
              astro            2      158.75       <.0001
              degree           4      939.71       <.0001
              uv               1      100.77       <.0001
```

or 22.6 for the reciprocal odds ratio corresponding to reversing category order of one of the variables; that is, for subjects with less than a high school degree the estimated odds of believing that astrology is very scientific (versus not scientific at all) are 22.6 times the estimated odds for those with a graduate degree. This odds ratio estimate also is the odds ratio using the corner-fitted values, $(22.01 \times 147.40)/(104.60 \times 1.37) = 22.6$.

The 95% Wald confidence interval for the common local odds ratio e^{β} is $\exp[-0.390 \pm (1.96 \times 0.042)]$, or (0.62, 0.74). The corresponding 95% confidence interval for the odds ratio for the four corner cells is $\exp[9(-0.390) \pm 1.96 \times$

9(0.042)], or (0.023, 0.086). This is a strong association, perhaps more clearly portrayed by the corresponding confidence interval (11.7, 43.7) for the reciprocal corner-cells odds ratio.

For testing H_0: $\beta = 0$, the likelihood-ratio statistic $G^2(\text{I} \mid \text{L} \times \text{L}) = 107.6 - 6.8 = 100.8$. The Wald statistic $(\hat{\beta}/\text{SE})^2 = (-0.3898/0.0421)^2 = 85.7$. For these equally spaced scores, the score statistic $(n-1)\hat{\rho}^2 = 1792(0.2255)^2 = 91.1$. All three statistics show extremely strong evidence of an association (df $= 1$, P-value < 0.0001).

For scores $\{v_j = j\}$, the marginal mean and standard deviation for opinion about astrology are 2.63 and 0.58. The standardized scores are $\{(j - 2.63)/0.58\}$, or $(-2.81, -1.08, 0.65)$. Unit-spaced scores for highest degree have standard deviation 1.19, and the standardized equal-interval scores are $(-1.38, -0.54, 0.30, 1.14, 1.99)$. For these scores, the standardized association parameter estimate is $\hat{\beta}^* = -0.390(0.58)(1.19) = -0.267$. By solving $\hat{\beta}^* = \hat{\rho}^*/[1 - (\hat{\rho}^*)^2]$ for $\hat{\rho}^*$, we find $\hat{\rho}^* = -0.25$. If there is an underlying bivariate normal distribution for the variables, we estimate the correlation to be -0.25, a relatively weak negative association.

6.3 ROW OR COLUMN EFFECTS ASSOCIATION MODELS

Generalizations of the linear-by-linear association model treat scores for the categories of a variable as parameters rather than as fixed by the data analyst. This approach provides estimated scores for which the fit is best. The extra flexibility of not needing to choose the scores seems advantageous, but it comes at a price. When the scores are parameters, then unless we place an order restriction on the parameters, the results are invariant to the category ordering. That is, if we refit the model after permuting the order of the categories in some way, we'll get the same fit with similarly permuted values of the estimated scores. Thus, the variable is treated as nominal, and the model is less parsimonious. However, such a model with parameter scores for one of the variables is natural when one variable is nominal and one is ordinal.

6.3.1 Row Effects Model for a Nominal-Ordinal Association

Suppose that the row variable, X, is a nominal variable, and the column variable, Y, is ordinal. As in the L $\times$ L model, we represent the ordered columns with monotone scores, $v_1 \leq v_2 \leq \cdots \leq v_c$. Since the rows are unordered, we represent them by parameters instead of scores. Replacing the ordered values $\{\beta u_i\}$ in the linear-by-linear term $\beta u_i v_j$ in model (6.2) by unordered parameters $\{\mu_i\}$,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j. \tag{6.5}$$

Identifiability requires constraints such as $\lambda_r^X = \lambda_c^Y = \mu_r = 0$. This model has $r - 1$ more parameters (the $\{\mu_i\}$) than the independence model. Independence is the special case $\mu_1 = \cdots = \mu_r$.

Odds ratios describe the way that $\{\mu_i\}$ determine the association structure. The log odds ratio comparing rows a and b for adjacent columns j and $j + 1$ is

$$\log \frac{\mu_{aj}\mu_{b,j+1}}{\mu_{a,j+1}\mu_{bj}} = (\mu_b - \mu_a)(v_{j+1} - v_j).$$

For all j, such log odds ratios have the same sign as $\mu_b - \mu_a$. So this model implies that the rows are stochastically ordered on their conditional distributions of Y. The stochastic ordering is determined by the $\{\mu_i\}$. If $\mu_a > \mu_b$, observations in row a tend to be higher on Y than are observations in row b. The greater the difference between μ_a and μ_b, the greater the difference between the two conditional distributions. If $\mu_a = \mu_b$, the conditional distributions on Y are identical in rows a and b. For the unit-spaced scores $\{v_{j+1} - v_j = 1\}$, the log odds ratio equals $(\mu_b - \mu_a)$ for each pair of adjacent columns.

The $\{\mu_i\}$ are *row effects* and model (6.5) is called a *row effects model*. A corresponding *column effects model* has association term $u_i v_j$ that treats X as ordinal with fixed scores $\{u_i\}$ and Y as nominal with parameters $\{v_j\}$.

6.3.2 Corresponding Adjacent-Categories Logit Model

With $\{v_{j+1} - v_j = 1\}$, the row effects model has adjacent-categories logit form with proportional odds structure,

$$\log \frac{P(Y = j + 1 \mid x = i)}{P(Y = j \mid x = i)} = \alpha_j + \mu_i. \tag{6.6}$$

The effect in row i is identical for each pair of adjacent responses.

Plots of these adjacent-categories logits against i $(i = 1, \ldots, I)$ for different j are parallel. This is model (4.4) proposed by Simon (1974) described in Section 4.1.2, which Goodman (1983) called the *parallel odds model*.

6.3.3 Model Fitting and Inference for the Row Effects Model

The likelihood equations for the row effects model (6.5) are

$$\hat{\mu}_{i+} = n_{i+}, \quad i = i, \ldots, r, \qquad \hat{\mu}_{+j} = n_{+j}, \quad j = 1, \ldots, c,$$

and

$$\sum_j v_j \hat{\mu}_{ij} = \sum_j v_j n_{ij}, \qquad i = 1, \ldots, r.$$

Regarding Y as a response variable and letting $p_{j|i} = n_{ij}/n_{i+}$, since $\hat{\mu}_{i+} = n_{i+}$ the third likelihood equation can be expressed as

$$\sum_j v_j \hat{\pi}_{j|i} = \sum_j v_j p_{j|i}.$$

For the conditional distribution within each row, the mean score on Y is the same for the fitted distribution and the sample data. Although the $\{\mu_i\}$ are row parameters and not means, variability among $\{\mu_i\}$ represents variability among conditional means on Y in the population modeled. In fact, the ML estimates $\{\hat{\mu}_i\}$ necessarily have the same ordering as the the sample row means $\{M_i = \sum_j v_j p_{j|i}\}$ (Agresti et al. 1987a).

The likelihood equations are solved using iterative methods such as the Newton–Raphson method. Since the model is loglinear, it is easily fitted with software for generalized linear models (see the Appendix). The G^2 and X^2 goodness-of-fit statistics have large-sample chi-squared null distributions with df $= (r - 1)(c - 2)$.

Infinite estimates occur when the sufficient statistics $\{M_i = \sum_j v_j p_{j|i}\}$ for $\{\mu_i\}$ take their minimum or maximum possible values for the marginal counts observed. Specifically, $\hat{\mu}_i = -\infty$ if all observations in row i fall in the first column, and $\hat{\mu}_i = \infty$ if all observations fall in the last column.

6.3.4 Example: Happiness and Marital Status

Table 6.4 displays the relationship between marital status and happiness for data from the 2006 General Social Survey. Goodness-of-fit tests show that the model of independence (I) is inadequate, with deviance $G^2(\text{I}) = 218.0$ for df $= 4$ (P-value < 0.0001). The table also shows standardized residuals for the model. The count in the very happy category was much higher than independence predicts for those who are married and much lower than expected for the other two groups.

Denote the independence model by I and the row effects model by R. Table 6.5 shows output for the R model. Adding the row effects parameters dramatically improves the fit $[G^2(\text{R}) = 1.3, \text{df} = 2]$ compared to the I model. Testing independence (H_0: $\mu_1 = \mu_2 = \mu_3$) using $G^2(\text{I} \mid \text{R}) = G^2(\text{I}) - G^2(\text{R}) = 216.7$ (df $= 2$) provides extremely strong evidence of an association. Table 6.4 also shows fitted values for the two models. The improved fit is noticeable in every cell, but especially in the cells for the very happy category.

TABLE 6.4. Happiness and Marital Status, with Standardized Residuals for Independence Model in Parentheses[a]

	Happiness		
Marital Status	Not Too Happy	Pretty Happy	Very Happy
Married	93 (−9.6)	720 (−5.7)	600 (12.9)
	175.5*, 88.9[†]	794.3, 728.2	443.2, 595.9
Divorced or separated	119 (6.5)	355 (2.4)	112 (−7.2)
	72.8, 117.0	329.4, 358.9	183.8, 110.0
Never married	127 (4.8)	459 (4.2)	144 (−7.9)
	90.7, 133.0	410.3, 446.9	229.0, 150.0

Source: 2006 General Social Survey.

[a] The first entry (*) in each pair is the fit of the independence model; the second entry (†) is the fit of the row effects model.

TABLE 6.5. Output for Fitting Loglinear Row Effects Model to Table 6.4

Parameter		DF	Estimate	Standard Error	Like. Ratio 95% Confidence Limits		Wald Chi-Square
Intercept		1	5.0109	0.0733	4.8645	5.1518	4674.12
marital	1	1	-1.2939	0.1724	-1.6333	-0.9573	56.33
marital	2	1	-0.0373	0.1871	-0.4044	0.3292	0.04
marital	3	0	0.0000	0.0000	0.0000	0.0000	.
happy	1	1	-0.1202	0.1190	-0.3538	0.1129	1.02
happy	2	1	1.0914	0.0727	0.9503	1.2356	225.15
happy	3	0	0.0000	0.0000	0.0000	0.0000	.
v*marital	1	1	0.8910	0.0772	0.7407	1.0435	133.08
v*marital	2	1	-0.0910	0.0892	-0.2659	0.0837	1.04
v*marital	3	0	0.0000	0.0000	0.0000	0.0000	.

The output uses indicator variables for the first two categories of each classification. The interaction term equals the product of the score for happiness and a parameter for marital status. Thus, the row effect estimates satisfy $\hat{\mu}_3 = 0$, and the other two estimates contrast the married and the divorced/separated with those never married. The estimates are $\hat{\mu}_1 = 0.891$ and $\hat{\mu}_2 = -0.091$. The further $\hat{\mu}_i$ falls in the positive direction, the greater the tendency for the marital status i to locate at the very happy end of the happiness scale relative to those who have never been married. So those who are married have much more of a tendency to be very happy than the other two marital statuses, and those who are divorced/separated seem to be similar to those who have never been married.

Loglinear models do not distinguish between response and explanatory variables. To treat happiness as a response, we could use the equivalent adjacent-categories logit model. From (6.6) the model predicts constant odds ratios for adjacent columns of happiness. For example, since $\hat{\mu}_3 - \hat{\mu}_1 = 0.891$, the estimated odds that the married were very happy instead of pretty happy, or pretty happy instead of not too happy, were $\exp(0.891) = 2.4$ times the corresponding estimated odds for those never married. The 95% profile likelihood confidence interval for this odds ratio is $[\exp(0.7407), \exp(1.0435)] = (2.1, 2.8)$. The estimated odds that the married were very happy instead of not too happy were $\exp[2(0.891)] = 5.9$ times the corresponding estimated odds for those never married, with a 95% confidence interval (4.4, 8.1).

6.3.5 Order-Restricted Row or Column Effect Parameters

To treat a variable as ordinal, association models use fixed monotone scores. With parameter scores such as the $\{\mu_i\}$ in the row effects model, the resulting ML estimates of scores need not be monotone. Variables with parameter scores are treated as nominal.

Constrained versions of the models with parameter scores recognize the ordinality by maximizing the likelihood subject to order restrictions. For example, suppose that we want to treat the row variable as ordinal with the row effects model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j.$$

We might prefer this model over the linear-by-linear association model if we prefer not to assign scores to the rows. To treat the rows as ordinal but with unspecified scores, we could fit the model subject to the ordering constraint

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_r.$$

Let $\{\hat{\mu}_i^*\}$ denote the ML estimates of $\{\mu_i\}$ subject to the ordering constraint. Agresti et al. (1987a) noted that the ML estimates $\{\hat{\mu}_i\}$ for the ordinary row effects model satisfy the constraint $\hat{\mu}_1 \leq \hat{\mu}_2 \leq \cdots \leq \hat{\mu}_r$ if the sample conditional row means $M_i = \sum_j v_j p_{j|i}$ are monotone increasing. Otherwise, the order-restricted solution $\{\hat{\mu}_i^*\}$ corresponds to fitted values that have means $\{M_i^*\}$ such that $M_a^* = M_{a+1}^*$ whenever $\hat{\mu}_a^* = \hat{\mu}_{a+1}^*$. If the ordinary ML estimates $\{\hat{\mu}_i\}$ satisfy the order restriction except for $\hat{\mu}_a > \hat{\mu}_{a+1}$, then $\{\hat{\mu}_i^*\}$ can be determined by combining row a and $a+1$ and refitting the model. If the resulting $\{\hat{\mu}_i^*\}$ are in proper order, that is the final solution. Otherwise, out-of-order rows are again combined and the model is refitted, continuing in this way until the order restriction is satisfied. (This fitting method employs the *pooling adjacent violators* algorithm, presented in Section 7.5.1.)

Let $G^2(\mathrm{R}^*)$ denote the deviance for the fit of the order-restricted model to the original table. Let $G^2(\mathrm{R}')$ denote the fit of the row effects model to the collapsed table for which rows are combined that have identical $\hat{\mu}_i^*$ order-restricted estimates. In addition, let $G^2(\mathrm{I})$ and $G^2(\mathrm{I}')$ denote the deviances for the fit of the independence model to the original and collapsed tables, respectively. Then, Agresti et al. (1987a) showed that

$$G^2(\mathrm{R}^*) = G^2(\mathrm{R}') + [G^2(\mathrm{I}) - G^2(\mathrm{I}')] = G^2(\mathrm{R}') + \sum_k G^2(\mathrm{I}_k), \qquad (6.7)$$

where $G^2(\mathrm{I}_k)$ is the deviance for the fit of the independence model to the kth set of rows that are combined and that have identical order-restricted score parameter estimates.

Goodman (1985) presented tests about equality of score parameters for models with row effects or column effects. His test statistics relate to goodness-of-fit statistics for order-restricted solutions. For example, to test H_0: $\mu_i = \mu_{i+1}$ for the row effects model, Goodman proposed the test statistic

$$T = [G^2(\mathrm{I}) - G^2(\mathrm{R})] - [G^2(\mathrm{I}') - G^2(\mathrm{R}')],$$

where I' and R' refer to the collapsed table that combines rows i and $i+1$. Under H_0, T has an asymptotic chi-squared distribution with df $= 1$. Now if the order-restricted fit has strict inequality except for $\hat{\mu}_i^* = \hat{\mu}_{i+1}^*$, then

$$G^2(\mathrm{R}^*) = G^2(\mathrm{R}') + [G^2(\mathrm{I}) - G^2(\mathrm{I}')] = G^2(\mathrm{R}) + T,$$

so that

$$T = G^2(\mathrm{R}^*) - G^2(\mathrm{R}).$$

Suppose that the R model truly holds with strictly ordered parameter scores, $\mu_1 < \mu_2 < \cdots < \mu_r$. Then $G^2(R^*)$ has the same asymptotic distribution as $G^2(R)$: namely, chi-squared with df $= (r-1)(c-2)$. By contrast, suppose that $\mu_1 < \cdots < \mu_i = \mu_{i+1} < \cdots < \mu_r$. Then, with limiting probability 0.5 as $n \to \infty$, the ML estimates are in order and $G^2(R^*) = G^2(R)$ has an asymptotic chi-squared distribution with df $= (r-1)(c-2)$. Also, with limiting probability 0.5 as $n \to \infty$, $\hat{\mu}_i > \hat{\mu}_{i+1}$ and $G^2(R^*)$ has an asymptotic chi-squared distribution with df $= (r-1)(c-2) + 1$. Thus, in this case, the limiting distribution is a mixture of two chi-squared distributions, which is a special case of a *chi-bar-squared distribution* (Section 7.5.1).

6.3.6 Example: Boys' Disturbed Dreams Revisited

Section 4.3.7 used the stereotype model (4.15) introduced in Section 4.3 to analyze a 5×4 data set (Table 4.4) from a study that cross-classified boys by their age and by the severity of their disturbed dreams. The model used midpoint scores (6, 8.5, 10.5, 12.5, 14.5) for the five age levels and parameters for the four categories related to the severity of disturbed dreams. A related approach fits the column effects model, with parameters for the columns,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + u_i v_j.$$

This model has $G^2(C) = 9.75$ with df $= 9$. With constraint $v_4 = 0$, the ML estimates of the column score parameters are

$$\hat{v}_1 = 0.309, \quad \hat{v}_2 = 0.059, \quad \hat{v}_3 = 0.112, \quad \hat{v}_4 = 0.0.$$

These suggest a negative trend in the association, with dreams tending to be less severe at greater ages, except for the out-of-order pair $\hat{v}_2 < \hat{v}_3$.

Now suppose we assume that $v_1 \geq v_2 \geq v_3 \geq v_4$ and conjecture that $\hat{v}_2 < \hat{v}_3$ is due simply to sampling error. This implies that boys of higher ages are stochastically lower on severity of disturbed dreams. The order-restricted column effects model has

$$\hat{v}_1^* = 0.309, \quad \hat{v}_2^* = \hat{v}_3^* = 0.086, \quad \hat{v}_4^* = 0.0.$$

The order restricted fit has $G^2(C^*) = 10.13$, very nearly as good as the unrestricted fit. If truly $v_1 > v_2 = v_3 > v_4$, this fit statistic has asymptotic null distribution that is an equal mixture of chi-squared with df $= 9$ and df $= 10$, so the order-restricted model is plausible. The test of H_0: $v_2 = v_3$ has test statistic $T = G^2(C^*) - G^2(C) = 10.13 - 9.75 = 0.38$, with df $= 1$. However, it is best to treat T as only an informal index, since H_0 was suggested by the fit of the unrestricted model. The statistic does support the conjecture that the order violation may merely reflect sampling error.

For this table the independence model fits poorly, having $G^2(I) = 32.46$ with df $= 12$. The 5×3 table that combines columns 2 and 3 of the original table has

$G^2(I') = 30.97$ and $G^2(C') = 8.64$. The decomposition (6.7) in the context of column effects is

$$G^2(C^*) = 10.13 = G^2(C') + [G^2(I) - G^2(I')] = 8.64 + 32.46 - 30.97.$$

6.3.7 Row or Column Effects Association as a Stereotype Model

As mentioned, in Section 4.3.7 we analyzed Table 4.4 using the stereotype model. That model treated the severity of disturbed dreams as the response variable. In fact, the column effects model fitted above is equivalent to the special case of the stereotype model,

$$\log \frac{\pi_j}{\pi_4} = \alpha_j + \phi_j \beta u_i, \qquad j = 1, 2, 3.$$

The parameter v_j in the column effects model corresponds to $\phi_j \beta$ in the stereotype model. For the constraints $\phi_1 = 1.0$ and $\phi_4 = 0.0$ for the stereotype model, $\beta = v_1$. The order-restricted column effects model is equivalent to the ordered stereotype model of Section 4.3.3.

Alternatively, for these data we could use the simpler linear-by-linear association model with the same age scores and with equally spaced scores for severity. That model corresponds to the stereotype model with equally spaced $\{\phi_j\}$ and the same age scores. It has fit statistic $G^2 = 14.61$, which is 4.86 higher (with two fewer parameters) than the unconstrained column effects model. The L × L estimate $\hat{\beta} = -0.097$ (SE = 0.024) also shows the overall negative trend in the association. In yet another analysis, Gelfand and Kuo (1991) assumed stochastically ordered distributions on disturbed dreams at various age levels.

6.4 ASSOCIATION MODELS FOR MULTIWAY TABLES

We next study ways to describe association in multiway contingency tables. Multiway tables with ordinal responses can use generalizations of the association models presented in the two preceding sections.

6.4.1 Homogeneous Conditional Associations

Consider a three-way contingency table $\{n_{ijk}\}$ with expected frequencies $\{\mu_{ijk}\}$. The useful loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \tag{6.8}$$

permits a conditional association between each pair of variables. It has the structure of *homogeneous association*, whereby a particular odds ratio between two variables is the same at each category of the third variable. This model is often denoted by (XY, XZ, YZ), representing the fact that its minimal sufficient statistics are the

three sets of two-way marginal tables collapsed over the third variable. The model (XZ, YZ) of conditional independence between X and Y, given Z, is the special case in which $\{\lambda_{ij}^{XY} = 0\}$. The more general model with a three-factor interaction term λ_{ijk}^{XYZ} allows each conditional odds ratio to vary across categories of the third variable and is saturated.

These loglinear models treat all the variables as nominal. Analogous models apply for higher-dimensional tables. By contrast, the rich collection of ordinal models includes association models that are more parsimonious than the model of homogeneous association and models permitting heterogeneous association that are unsaturated.

6.4.2 Modeling Ordinal Conditional Associations

In a three-way table, suppose that X and Y are ordinal. Models for the XY conditional association that are special cases of the homogeneous association model (XY, XZ, YZ) replace the λ_{ij}^{XY} association term by a structured term that accounts for ordinality. The linear-by-linear term $\beta u_i v_j$, row effects term $\mu_i v_j$, and column effects term $u_i v_j$, respectively, provide a stochastic ordering of conditional distributions within rows and within columns, within rows, and within columns.

With a linear-by-linear term, the model is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{6.9}$$

The conditional local odds ratios $\{\theta_{ij|k}^L\}$ describing local XY association at a given value of Z [see (2.15) for sample values] then satisfy

$$\log \theta_{ij|k}^L = \beta(u_{i+1} - u_i)(v_{j+1} - v_j) \qquad \text{for all } k.$$

The association is the same in different partial tables, with *homogeneous linear-by-linear XY association*. For equally spaced scores for X and for Y, there is *homogeneous uniform XY association*. For example, with unit-spaced scores, e^β is the common value of all $(r-1)(c-1)$ local odds ratios in each partial table. This model has one more parameter than the model (XZ, YZ) of XY conditional independence. With K categories for Z, its residual df $= K(r-1)(c-1) - 1$.

When the association is heterogeneous, structured terms for ordinal variables make effects simpler to interpret than in the saturated model. For example, consider the *heterogeneous linear-by-linear XY association model*

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + (\beta + \beta_k)u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \tag{6.10}$$

with a constraint such as $\beta_K = 0$. Such a model has linear-by-linear association within each level of Z, but the strength of the XY association varies across levels of Z. With unit-spaced scores, the XY conditional local odds ratios satisfy

$$\log \theta_{ij|k}^L = \beta + \beta_k \qquad \text{for all } i \text{ and } j.$$

Here β represents the uniform local log odds ratio at level K of Z, and β_k represents the difference between the uniform local log odds ratio at levels k and K. Alternatively, we could parameterize the model by replacing $\beta + \beta_k$ by β_k. Then no constraint is needed and β_k is the uniform local log odds ratio in level k. Fitting this model is equivalent to fitting the L × L model (6.2) separately at each level of Z. With K strata, it has K more parameters than the model (XZ, YZ) of XY conditional independence. Its residual df $= K[(r-1)(c-1) - 1]$.

The models also generalize to describe association structure among several response variables, some of which are ordinal and some of which are nominal. A basic homogeneous association model uses a term of linear-by-linear form $\beta u_i v_j$ for a pair of ordinal variables, a term of row-effects form $\mu_i v_j$ for an association between a nominal variable and an ordinal variable (with parameters $\{\mu_i\}$ for the nominal variable), and a term of form λ_{ij} for an association between two nominal variables.

6.4.3 Example: Income and Education by Race

Table 6.6, from the 2006 General Social Survey, shows the relationship between $I =$ family income and $E =$ education in the United States separately for black and white categories of race (R). The loglinear model (IR, ER) of conditional independence between income and education, given race, fits poorly $(G^2 = 287.2,$ df $= 8)$.

Table 6.7 shows the results for several models that permit conditional association. The homogeneous linear-by-linear EI association model using equally spaced scores fits much better, with only one additional parameter. However, it shows some lack of fit $(G^2 = 23.0,$ df $= 7)$. The heterogeneous linear-by-linear EI association model fits only a bit better $(G^2 = 19.9,$ df $= 6)$, with a stronger estimated local log odds ratio for black subjects (0.840) than for white subjects (0.577). However, it is also worth considering other models that have a lack of three-factor interaction. The homogeneous row effects model, using parameter scores $\{\mu_i\}$ for education levels, has $G^2 = 18.6$ (df $= 6$). The homogeneous column effects model, using parameter scores $\{v_j\}$ for income levels, has $G^2 = 4.3$ (df $= 6$).

The homogeneous column effects model gives an excellent fit, the reduction in deviance compared to the conditional independence model being 282.8 at the cost

TABLE 6.6. Educational Degree and Family Income, by Race

Highest Degree	Family Income, Black			Family Income, White		
	Below Average	Average	Above Average	Below Average	Average	Above Average
< High school	43	36	5	114	97	12
High school, junior college	104	140	23	410	658	221
College, grad school	16	30	18	97	259	287

Source: 2006 General Social Survey.

TABLE 6.7. Goodness-of-Fit of Loglinear Models Fitted to Table 6.6 on Education (E), Income (I), and Race (R)

Model	$\log \mu_{ijk}$	G^2	df
EI cond. indep.	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	287.2	8
EI homo. L × L	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + \beta u_i v_j + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	23.0	7
I effects homo.	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + u_i v_j + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	4.3	6
E effects homo.	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + \mu_i v_j + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	18.6	6
EI hetero. L × L	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + \beta_k u_i v_j + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	19.9	6
Homo. assoc.	$\lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + \lambda_{ij}^{EI} + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}$	2.8	4

of only two extra parameters. This model is

$$\log \mu_{ijk} = \lambda + \lambda_i^E + \lambda_j^I + \lambda_k^R + u_i v_j + \lambda_{ik}^{ER} + \lambda_{jk}^{IR}.$$

With education scores $\{u_i = i\}$ and constraint $\hat{v}_3 = 0$, the column effect estimates are $\hat{v}_1 = -1.668$ (SE = 0.107) and $\hat{v}_2 = -1.137$ (SE = 0.095). Since $\{u_i\}$ are monotone increasing, the monotone increase in $\{\hat{v}_j\}$ reflects the positive association between education and income, controlling for race. The estimated local log odds ratios are about half as large for columns 1 and 2 as they are for columns 2 and 3. The estimated odds that income is above average instead of average are multiplied by $\exp(1.137) = 3.1$ for each category increase in education. By contrast, the estimated odds that income is average instead of below average are multiplied by $\exp(1.668 - 1.137) = 1.7$ for each category increase in education. The estimated odds ratio for the four corner cells in each partial table is $\exp[(u_3 - u_1)(\hat{v}_3 - \hat{v}_1)] = \exp[2(1.668)] = 28.1$, reflecting a very strong conditional EI association.

6.4.4 Testing Conditional Independence in Three-Way Tables

In many studies, a major focus is whether the association between two variables disappears after controlling for other variables. In testing the hypothesis of conditional independence between X and Y when one or both are ordinal, models can use terms that reflect an expected ordinal feature such as trend. We then test whether a model parameter equals 0 under which conditional independence occurs. The test statistic could be a likelihood-ratio statistic, Wald statistic, or score statistic. When such models fit relatively well, the tests are more powerful than tests that ignore the ordinality, such as tests comparing models (XZ, YZ) and (XY, XZ, YZ). In Section 3.7.3 we presented such tests using cumulative logit models. Alternatively, we could use association models as the basis for such tests.

When both X and Y are ordinal, we can compare the conditional independence model (XZ, YZ) to the homogeneous linear-by-linear XY association model (6.9). Tests of $\beta = 0$ in that model have df = 1. To allow the conditional association to vary among the strata, we could, instead, compare the conditional independence

model to the heterogeneous linear-by-linear association model, (6.10), the test having $df = K$. Unless the association truly varies substantially, being positive for some strata and negative for others, this test tends to be less powerful than the test of $\beta = 0$ for model (6.9) because of its greater df value.

When X is nominal and Y is ordinal, we could assign scores $\{v_j\}$ to Y and fit the model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \mu_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{6.11}$$

This homogeneous row effects model uses the same $\{\mu_i\}$ in each stratum. Conditional independence is H_0: $\mu_1 = \cdots = \mu_r$. The asymptotic chi-squared distribution has $df = r - 1$.

6.4.5   Generalized Cochran–Mantel–Haenszel Tests for $r \times c \times K$ Tables

With binary X and Y, a popular test of XY conditional independence for $2 \times 2 \times K$ contingency tables $\{n_{ijk}\}$ is the *Cochran–Mantel–Haenszel* (CMH) *test*. The test statistic sums over the strata the difference between n_{11k} and its expected value under the null hypothesis of conditional independence, conditional on the marginal totals for each stratum.

Mantel (1963) generalized the CMH statistic for ordinal X and Y. With ordered scores $\{u_i\}$ and $\{v_j\}$, it is sensitive to a correlation of common sign in each stratum. Evidence of a positive trend occurs if in each stratum $T_k = \sum_i \sum_j u_i v_j n_{ijk}$ exceeds its null expectation. Given the marginal totals in each stratum, under conditional independence

$$E(T_k) = \frac{\left(\sum_i u_i n_{i+k} \right) \left(\sum_j v_j n_{+jk} \right)}{n_{++k}},$$

$$\text{Var}\ (T_k) = \frac{1}{n_{++k} - 1} \left[\sum_i u_i^2 n_{i+k} - \frac{\left(\sum_i u_i n_{i+k} \right)^2}{n_{++k}} \right]$$

$$\left[\sum_j v_j^2 n_{+jk} - \frac{\left(\sum_j v_j n_{+jk} \right)^2}{n_{++k}} \right].$$

The statistic $[T_k - E(T_k)]/[\text{Var}(T_k)]^{1/2}$ equals $\sqrt{n_{++k} - 1}\,\hat{\rho}_k$ for the sample correlation $\hat{\rho}_k$ between X and Y in stratum k. To summarize across the K strata, Mantel proposed

$$M^2 = \frac{\left\{ \sum_k \left[\sum_i \sum_j u_i v_j n_{ijk} - E\left(\sum_i \sum_j u_i v_j n_{ijk} \right) \right] \right\}^2}{\sum_k \text{Var}\left(\sum_i \sum_j u_i v_j n_{ijk} \right)}. \tag{6.12}$$

This has an approximate χ_1^2 null distribution [see also Birch (1965)].

Landis et al. (1978) further generalized the CMH statistic to include M^2 and statistics for which one or both variables can be nominal. The tests treat X and Y symmetrically. Conditional on row and column totals, each stratum has $(r-1)(c-1)$ nonredundant cell counts. Let

$$\mathbf{n}_k = (n_{11k}, n_{12k}, \ldots, n_{1,c-1,k}, \ldots, n_{r-1,c-1,k})'.$$

Let $\boldsymbol{\mu}_k = E(\mathbf{n}_k)$ under H_0: conditional independence: namely,

$$\boldsymbol{\mu}_k = (n_{1+k}n_{+1k}, n_{1+k}n_{+2k}, \ldots, n_{r-1,+,k}n_{+,c-1,k})'/n_{++k}.$$

Let $\mathbf{V}_k$ denote the null covariance matrix of $\mathbf{n}_k$, where

$$\text{Cov}\,(n_{ijk}, n_{i'j'k}) = \frac{n_{i+k}(\delta_{ii'}n_{++k} - n_{i'+k})n_{+jk}(\delta_{jj'}n_{++k} - n_{+j'k})}{n_{++k}^2(n_{++k} - 1)}$$

with $\delta_{ab} = 1$ when $a = b$ and $\delta_{ab} = 0$ otherwise. Let $\mathbf{B}_k = \mathbf{u}_k \otimes \mathbf{v}_k$ denote the Kronecker product matrix of constants based on row scores $\mathbf{u}_k$ and column scores $\mathbf{v}_k$ for stratum k. The generalized statistic is

$$L^2 = \left[\sum_k \mathbf{B}_k(\mathbf{n}_k - \boldsymbol{\mu}_k)\right]' \left[\sum_k \mathbf{B}_k\mathbf{V}_k\mathbf{B}_k'\right]^{-1} \sum_k \mathbf{B}_k(\mathbf{n}_k - \boldsymbol{\mu}_k).$$

Suppose that $\mathbf{u}_k = (u_1, \ldots, u_r)$ and $\mathbf{v}_k = (v_1, \ldots, v_c)$ for all strata. Then $L^2 = M^2$. Suppose that $\mathbf{u}_k$ is an $(r-1) \times I$ matrix $(\mathbf{I}, -\mathbf{1})$, where $\mathbf{I}$ is an identity matrix of size $(r-1)$ and $\mathbf{1}$ denotes a column vector of $r-1$ ones, and $\mathbf{v}_k = (v_1, \ldots, v_c)$. Then L^2 sums over the strata information about how r row means compare to their null expected values, and it has df $= r - 1$. That statistic treats X as nominal and Y as ordinal. Rank score versions of these statistics are analogs for ordered categorical responses of strata-adjusted Spearman correlation and Kruskal–Wallis tests. The generalized CMH tests are available in some software, such as PROC FREQ in SAS (see the Appendix).

6.4.6 CMH Tests and Related Score Tests for Models

The generalized CMH tests seem to be non-model-based alternatives to tests using models. However, a close connection exists between them. For various models, the generalized CMH tests are score tests about XY association parameters, just as the ordinary CMH test for $2 \times 2 \times K$ tables is a score test for the homogeneous association loglinear model and the equivalent logistic model for Y having main effects for X and for Z.

Consider first stratified ordinal-by-ordinal tables. Mantel's M^2 statistic based on $\sum_k(\sum_i \sum_j u_i v_j n_{ijk})$ for row and column scores [see formula (6.12)] is the score test statistic for testing H_0: $\beta = 0$ in the homogeneous linear-by-linear association

model (6.9) using the same scores. The statistic $\sum_k (\sum_i \sum_j u_i v_j n_{ijk})$ is the sufficient statistic for β in that model. With equally spaced scores for $\{v_j\}$, it is the score statistic for testing H_0: $\beta = 0$ in the adjacent-categories logit model

$$\log \frac{P(Y = j + 1)}{P(Y = j)} = \alpha_{jk} + \beta u_i,$$

which treats Y alone as a response variable and assumes the same linear effect of X on Y in each partial table. For the corresponding cumulative logit model,

$$\text{logit } [P(Y \leq j)] = \alpha_{jk} + \beta u_i,$$

the score statistic for testing H_0: $\beta = 0$ is Mantel's M^2 with $\{v_j\}$ scores that are midrank (or ridit) scores for Y.

Next consider stratified nominal-ordinal tables. The L^2 statistic mentioned for this case in Section 6.4.5 is a score statistic for testing H_0: $\mu_1 = \cdots = \mu_r$ in the loglinear homogeneous row effects model (6.4.5) having term $\mu_i v_j$ for the conditional XY association. With midrank $\{v_j\}$ scores, it is the score test for the cumulative logit model above that replaces βu_i by μ_i.

With large samples in each stratum, the generalized CMH tests give similar results as likelihood-ratio tests comparing the relevant model to the conditional independence model. An advantage of the model-based approach is that it provides estimates of effects as a by-product. Ultimately, this is more important than mere significance testing. However, the CMH tests have the advantage that they are also valid for randomized studies in which the samples cannot meaningfully be regarded as multinomialsamples (e.g., volunteer samples).

6.4.7 Example: Income and Education by Race, Revisited

We return to the analysis of Table 6.6 (Section 6.4.3) on $Y = $ family income and $X = $ education by $Z = $ race. Table 6.8 shows output from conducting some generalized CMH tests. Statistics treating a variable as ordinal used scores (1, 2, 3) for its categories.

The test with *general association* alternative treats X and Y as nominal. It is sensitive to any association that is similar in each level of Z and is the score test for the loglinear model (XY, XZ, YZ). The test with *row mean scores differ* alternative treats rows as nominal and columns as ordinal. It is sensitive to variation among the row mean scores on Y when that variation is similar in each level of Z, and it is the score test for the homogeneous row effects model. The test for a corresponding *column mean scores differ* alternative would result from applying this test after interchanging rows with columns and would treat rows as ordinal and columns as nominal. It is the score statistic for the homogeneous column effects model that was seen to fit well in Section 6.4.3. Not shown in Table 6.8 is the fact that it has test statistic equal to 266.80 with df $= 2$. Finally, the test for the *nonzero correlation* alternative treats X and Y as ordinal and uses Mantel's statistic (6.12).

TABLE 6.8. Output for Generalized Cochran–Mantel–Haenszel Tests with Data from Table 6.6 on Education and Income, Controlling for Race

```
            Summary Statistics for income by educ
                    Controlling for race

     Cochran-Mantel-Haenszel Statistics (Based on Table Scores)
     Statistic    Alternative Hypothesis    DF      Value      Prob
--------------------------------------------------------------------
         1        Nonzero Correlation        1     251.4861    <.0001
         2        Row Mean Scores Differ      2     257.7836    <.0001
         3        General Association         4     291.1257    <.0001
```

It is sensitive to a similar linear trend in each level of Z and is the score test for the homogeneous linear-by-linear association model.

For this example, all statistics show strong evidence of a conditional association. The nonzero correlation alternative has the advantage of focusing the effect on a single degree of freedom. Even when the homogeneous linear-by-linear association shows some lack of fit, as it does for these data, this statistic is typically very powerful whenever the actual conditional association has somewhat of a monotone trend with the same direction in each partial table.

6.5 MULTIPLICATIVE ASSOCIATION AND CORRELATION MODELS

The linear-by-linear association (L × L) model is the special case of the row effects (R) model, in which the row parameter scores $\{\mu_i\}$ are replaced by fixed values $\{u_i\}$, and the special case of the column effects (C) model, in which the column parameter scores $\{v_j\}$ are replaced by fixed values $\{v_j\}$. The R and C models are themselves special cases of a more general model that has parameters for both the row *and* column scores.

6.5.1 RC Model

Replacing $\{u_i\}$ and $\{v_j\}$ in the L × L model (6.2) by parameters $\{\mu_i\}$ and $\{v_j\}$ yields the *row and column effects* (RC) model (Goodman 1979a),

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta \mu_i v_j. \tag{6.13}$$

Identifiability requires location and scale constraints on $\{\mu_i\}$ and $\{v_j\}$. The standardized constraints

$$\sum \mu_i \pi_{i+} = \sum v_j \pi_{+j} = 0,$$

$$\sum \mu_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1$$

correspond to means of 0 and standard deviations of 1 for each margin. We need yet another restriction, because if the model holds, it also holds if signs are reversed for β and the $\{\mu_i\}$, or for β and the $\{\nu_j\}$, or for the $\{\mu_i\}$ and $\{\nu_j\}$. If the scores are monotone, we prefer the choice of signs for which

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_r \quad \text{and} \quad \nu_1 \leq \nu_2 \leq \cdots \leq \nu_c.$$

The RC model is log-multiplicative, *not* loglinear, because the predictor is a multiplicative (rather than linear) function of parameters μ_i and ν_j. It treats classifications as nominal, because the same fit results from a permutation of rows or of columns. Parameter interpretation is simplest when at least one variable is ordinal, through the local log odds ratios

$$\log \theta_{ij} = \beta (\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j).$$

One use of the model is to generate a hypothetical order of categories, corresponding to the order of the estimated scores when the categories are partially ordered or the ordering is unknown. For example, Xie (1992) did this in studying social mobility for the categories (service class, routine nonmanual, petty bourgeoisie, farmer, skilled working class, semi- and unskilled working class, agricultural).

6.5.2 Model Fitting and Inference for the RC Model

When $\{\nu_j\}$ are fixed, the RC model simplifies to the row effects (R) model. When $\{\mu_i\}$ are fixed, the RC model simplifies to the column effects (C) model. Goodman (1979a) suggested an iterative model-fitting algorithm that exploits this. A cycle of the algorithm has two steps. For some initial guess of $\{\nu_j\}$, it estimates $\{\mu_i\}$ as in the R model. Then, treating the estimated $\{\mu_i\}$ from the first step as fixed, it estimates $\{\nu_j\}$ as in the C model. Those estimates serve as fixed column scores in the first step of the next cycle, for reestimating $\{\mu_i\}$ in the R model. There is no guarantee of eventual convergence to the ML estimates, but this seems to happen when the model fits reasonably well.

This iterative procedure does not yield appropriate standard errors for the ML estimates, because each part of a cycle recognizes only a subset of the parameters as truly being parameters. Haberman (1995) proposed a stabilized Newton–Raphson algorithm for fitting association models that also generates appropriate standard errors and tends to converge even when the model fits poorly. Aït-Sidi-Allal et al. (2004) proposed a Fisher scoring algorithm. Specialized software exists that can fit such nonlinear models and provide standard errors (see the Appendix). The residual df value for testing the fit of the model is

$$\text{df} = rc - [1 + (r-1) + (c-1) + 1 + (r-2) + (c-2)] = (r-2)(c-2).$$

The model is of use only when $r > 2$ and $c > 2$; otherwise, it is saturated.

Although it may seem appealing to use parameter scores instead of arbitrary fixed scores, the RC model presents complications that do not occur with log-linear models. The likelihood may not be concave and may have local maxima. Ordinary first-order inference for the association parameter β may be inadequate (Simonoff and Tsai 1991). Independence is a special case, but it is awkward to test independence using the RC model. This is because of the lack of identifiability of the parameters under that condition. Independence can be characterized as $\beta = 0$, or as $\mu_1 = \cdots = \mu_r$, or as $\nu_1 = \cdots = \nu_c$. Because of this, Haberman (1981) showed that the asymptotic null distribution of $G^2(I) - G^2(RC)$ is not chi-squared. Instead, it is the same as that of $n\hat{\rho}^2$, where $\hat{\rho}$ is the canonical correlation for the table, which is the maximum correlation between the variables out of all ways of assigning scores to the rows and columns (monotone or not). That is, when the variables are independent, $[(G^2(I) - G^2(RC)) - n\hat{\rho}^2]$ converges in probability to 0. The null asymptotic distribution of either statistic is the same as that of the maximum eigenvalue of the $(r - 1) \times (r - 1)$ central Wishart matrix with $c - 1$ degrees of freedom.

The usual chi-squared asymptotic distributions *do* apply when we use likelihood-ratio statistics to compare the RC model to the simpler R model or C model or $L \times L$ model. The parameters are then identifiable under the special cases. For example, the R model is the special case of the RC model that replaces $\beta\nu_j$ by the fixed score ν_j.

6.5.3 Example: Mental Health Status and SES

Table 6.9 describes the relationship between children's mental health status and parents' socioeconomic status for a sample of 1660 residents of Manhattan. In introducing association models, Goodman (1979a, 1985) fitted various models to these data.

The RC model fits well ($G^2 = 3.6$, df $= 8$). For scaling (6.17), the ML estimates are $(-1.11, -1.12, -0.37, 0.03, 1.01, 1.82)$ for the row scores, $(-1.68, -0.14, 0.14, 1.41)$ for the column scores, and $\hat{\beta} = 0.17$. Nearly all estimated local log

TABLE 6.9. Mental Health and Parents' Socioeconomic Status

SES[a]	Mental Health			
	Well	Mild Symptoms	Moderate Symptoms	Impaired
A	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F	21	71	54	71

Source: Srole (1978), with permission.
[a] A to F, highest to lowest.

odds ratios are positive, as only the first two row scores are slightly out of order. This indicates a tendency for mental health to be better at higher levels of parents' SES.

Ordinal loglinear models also fit well. For equal-interval scores, $G^2(\text{L} \times \text{L}) = 9.9$ (df = 14). The statistic $G^2(\text{L} \times \text{L} \mid \text{RC}) = 6.3$ (df = 6) tests that row and column scores in the RC model are of equal-interval type. The parameter scores do not provide a significantly better fit. It is sufficient to use a uniform local odds ratio to describe the table. For unit-spaced scores, $\hat{\beta} = 0.091$ (SE = 0.015), so the fitted local odds ratio is $\exp(0.091) = 1.09$. There is strong evidence of positive association, but the degree of local association is rather weak. For the four corner cells, the estimated odds ratio is $\exp[(6 - 1)(4 - 1)0.091] = 3.9$, considerably stronger.

6.5.4 RC Model as a Special Case of the Stereotype Model

For the RC model (6.13), suppose that we treat Y as a response and X as explanatory. Forming baseline-category logits, we obtain

$$\log \frac{P(Y = j \mid X = i)}{P(Y = c \mid X = i)} = \log \frac{\mu_{ij}}{\mu_{ic}} = (\lambda_j^Y - \lambda_c^Y) + \beta \mu_i (\nu_j - \nu_c).$$

This has the form
$$\log \frac{P(Y = j \mid X = i)}{P(Y = c \mid X = i)} = \alpha_j + \phi_j \beta_i,$$

where $\alpha_j = \lambda_j^Y - \lambda_c^Y$, $\beta_i = \beta \mu_i$, and $\phi_j = \nu_j - \nu_c$. Now let $\mathbf{x}$ denote a vector of indicator variables for the rows, where $x_i = 1$ for row i and $x_i = 0$ otherwise, $i = 1, \ldots, r - 1$. Then the right-hand side of this equation has the form of the predictor $\alpha_j + \phi_j \boldsymbol{\beta}' \mathbf{x}$ of the *stereotype model* (4.15) presented in Section 4.3.

It follows that the RC model is a special case of the stereotype model. In Section 4.3.6 we presented the two-way contingency table version of that model. This connection is useful for both model interpretation and model fitting.

6.5.5 Latent Variable Model Implies the RC Model

Anderson and Vermunt (2000) showed that the RC model and more general association models result from latent variable models. Those underlying models related to graphical models for discrete and continuous variables (Lauritzen and Wermuth 1989).

The simplest case has the following structure: A latent variable Z is assumed such that the observed variables X and Y are conditionally independent given Z. Collapsed over Z, X and Y are assumed to have a multinomial joint distribution. Conditional on $X = i$ and $Y = j$, Z is assumed to have a normal distribution with mean $\mu(i, j)$ and variance σ^2. Under this structure, the distribution of X and Y has the RC form, with association term proportional to σ^2. The scores in the RC model then depend on the association between Z and (X, Y). For related results, see de Falguerolles et al. (1995).

Anderson and Vermunt also considered multivariate categorical responses. The same sort of latent structure then yields extensions of the RC model for multiple responses. Anderson (2002) extended graphical latent variable models to situations in which dependencies between observed variables are not fully accounted for by the latent variables.

6.5.6 Inequality Constraints on Model Parameters

In some applications it is reasonable to assume ordered scores but without specifying values for those scores as we did in Section 6.3.5 for the row effects model. To treat both rows and columns as ordinal with the *RC* model, we could fit the model subject to the ordering constraints

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_r \quad \text{and} \quad \nu_1 \leq \nu_2 \leq \cdots \leq \nu_c.$$

Ritov and Gilula (1991), Bartolucci and Forcina (2002), and Galindo-Garre and Vermunt (2004) gave algorithms for finding the order-restricted ML estimates. The X^2 test comparing the order-restricted model to the unrestricted RC model has a null distribution that is a mixture of chi-squared distributions that requires Monte Carlo or bootstrapping to estimate. In the special case that the order-restricted model has only two adjacent categories of one of the variables that have equal scores and all other scores are strictly monotone, the *P*-value is half that of a chi-squared random variable with df $= 1$.

For an order restriction such as $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_r$, the Bayesian approach can ensure that the posterior distribution recognizes this constraint by specifying the prior distribution in terms of parameters such as $\log(\mu_{j+1} - \mu_j)$ or by using densities over the nonnegative real line for $\{\phi_j = \mu_{j+1} - \mu_j\}$. When sample estimates violate the order, ML estimates commonly fall on the boundary of the constrained parameter space. An appealing aspect of the Bayesian approach is that posterior means of parameters often fall in the interior of the parameter space. Gelfand et al. (1992) and Iliopoulos et al. (2007, 2009) presented Bayesian fitting of parameter-constrained models, with the latter articles considering the RC model.

Although models with inequality constraints have the advantage of greater generality compared to models assuming a linear trend with fixed scores, a loss of power can result from the addition of parameters (e.g., the linear trend models have only a single association parameter). The model becomes less parsimonious, and tests of effects may be less powerful. Also, the user can no longer rely on standard chi-squared tests and confidence intervals for the effects, and the models are not easily available with standard software packages.

6.5.7 Correlation Models

The row and column effects (RC) model and its special cases use row and column scores or parameters to describe how the log cell probabilities differ from independence. Gilula (1984) and Goodman (1985) presented an alternative type of model

that does the same thing directly for the cell probabilities $\{\pi_{ij}\}$. In its simplest form with fixed row and column scores, this model is

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \rho u_i v_j). \tag{6.14}$$

When $\{u_i\}$ and $\{v_j\}$ are standardized to satisfy

$$\sum u_i \pi_{i+} = \sum v_j \pi_{+j} = 0 \quad \text{and} \quad \sum u_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1,$$

ρ is the correlation between the scores for joint distribution $\{\pi_{ij}\}$. That is, for these standardized scores, it follows directly from (6.14) that

$$\sum_i \sum_j u_i v_j \pi_{ij} = \rho.$$

Because of this, model (6.14) is called a *correlation model*. The independence model is the special case in which $\rho = 0$.

With fixed monotone scores, the model treats both variables as ordinal. This correlation model is then like the linear-by-linear association model in that it uses a single parameter to describe departures from independence. Thus, it has residual df $= rc - r - c$. For the standardized scores,

$$\sum_i u_i \left(\frac{\pi_{ij}}{\pi_{+j}} \right) = \rho v_j \quad \text{and} \quad \sum_j v_j \left(\frac{\pi_{ij}}{\pi_{i+}} \right) = \rho u_i$$

(Goodman 1985). Consider the second equation. Note that $\sum_j v_j(\pi_{ij}/\pi_{i+})$ is the mean response in row i, using the conditional distribution of Y given x in that row. So the equation implies that the mean response model of Section 5.6 holds, treating Y as a response variable with slope ρ. Similarly, the first equation states that the mean response model holds, treating X as a response variable, with slope ρ. When the model has ML estimate $\hat{\rho}$ close to zero, since

$$1 + \hat{\rho}u_i v_j \sim \exp(\hat{\rho}u_i v_j),$$

$\hat{\rho}$ is similar to $\hat{\beta}$ for the L $\times$ L model using the same scores.

A more general model lets the row and column scores be parameters,

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \rho\mu_i v_j). \tag{6.15}$$

In that case, the correlation model is also called the *canonical correlation model*, because ML estimates of the scores maximize the correlation for (6.15). As with the RC model, although parameter scores provide greater flexibility, the model does not then utilize the ordering of the categories.

A strand of related research, beginning with Karl Pearson, deals with estimating the correlation for an assumed underlying bivariate normal distribution. See Section 7.2.2 and Note 7.3.

6.5.8 Example: Mental Health and SES Revisited

In Section 6.5.3 we used association models to analyze the mental health data in Table 6.9. Goodman (1985) also fitted correlation models to these data. The canonical correlation model (6.15) with standardized parameter scores has estimated scores $(-1.09, -1.17, -0.37, 0.05, 1.01, 1.80)$ for the rows and $(-1.60, -0.19, 0.09, 1.48)$ for the columns, each monotone except for the first two rows. The model fits well ($G^2 = 2.75$, df $= 8$). The quality of fit and the estimated scores are similar to those that we showed in Section 6.5.3 for the RC model with standardized parameter scores.

More parsimonious correlation models also fit these data well. Models that have fixed monotone scores recognize the ordering of categories. With equally spaced scores, the model (6.14) has $G^2 = 9.64$ (df $= 14$), similar to the fit of the L × L association model with the same scores ($G^2 = 9.90$, df $= 14$). With standardized scores, the correlation estimate is $\hat{\rho} = 0.16$.

All analyses of Table 6.9 have yielded similar conclusions about the association. They all neglect, however, the fact that mental health is a natural response variable. It may make more sense to use a model that recognizes this, such as a cumulative logit model. The proportional odds form of that model with a linear effect of SES for scores (6, 5, 4, 3, 2, 1) also fits well ($G^2 = 10.87$, df $= 14$). It has effect $\exp(\hat{\beta}) = \exp(0.167) = 1.18$ for the common cumulative odds ratio for pairs of adjacent levels of SES, compared to 1.09 for the common local odds ratio for the L × L model.

6.5.9 Similarities Between Correlation Models and Association Models

The correlation model (6.14) with fixed scores is analogous to the linear-by-linear association model (6.2). The correlation model (6.15) with parameter scores has many properties in parallel with the RC model (6.13). Here we summarize some such results from Gilula et al. (1988).

Given fixed values for the marginal distributions, the row and column scores, and the correlation between them, the joint distribution $\{\pi_{ij}\}$ specified by the RC model maximizes the *entropy*, $\sum_i \sum_j \pi_{ij} \log \pi_{ij}$. By contrast, the joint distribution specified by the correlation model (6.15) minimizes $\sum_i \sum_j \pi_{ij}^2 / \pi_{i+} \pi_{+j}$. An implication is that under such conditions, the RC joint distribution is the closest to the independence distribution according to the *Kullback–Leibler distance*,

$$\sum_i \sum_j \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}},$$

whereas the joint distribution for the correlation model is closest to the independence distribution according to the *Pearsonian distance*,

$$\sum_i \sum_j \frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}}.$$

Goodman (1985) noted that the Pearsonian distance for the correlation model equals ρ^2, the square of the correlation parameter in that model.

Under the RC model, the entropy is an increasing function of β. Under the correlation model, a summary measure

$$\sum_{j=1}^{k} \frac{\pi_{ij}}{\pi_{i+}} - \sum_{j=1}^{k} \frac{\pi_{i+1,j}}{\pi_{i+1,+}}$$

of the difference between adjacent conditional distributions of one variable given the other is an increasing function of ρ for given $k = 1, \ldots, c - 1$ and $i = 1, \ldots, r - 1$. When the row parameter scores and the column parameter scores are scaled to have means of 0 and standard deviations of 1, the association parameters β from the RC model (6.13) and ρ from the correlation model (6.15) satisfy

$$\beta = \sum_i \sum_j \pi_{i+} \mu_i \pi_{+j} \nu_j \log \theta_{ij}^{\mathrm{L}} \quad \text{and} \quad \rho = \sum_i \sum_j \pi_{i+} \mu_i \pi_{+j} \nu_j (\pi_{ij} - \pi_{i+} \pi_{+j})$$

for the local odds ratios $\{\theta_{ij}^{\mathrm{L}}\}$.

Goodman (1985) discussed advantages of association models over correlation models. The correlation model is not defined for all possible combinations of score values because of the constraint $0 \le \pi_{ij} \le 1$. Also, its ML fitted values do not have the same marginal totals as the observed data, and it does not generalize in a simple way to multiway tables.

6.5.10   More General Association and Correlation Models

For the saturated loglinear model for a two-way table, $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, Goodman (1985, 1996) expressed the association term in a form that generalizes the $\beta \mu_i \nu_j$ term in the RC model. His expression is

$$\lambda_{ij}^{XY} = \sum_{k=1}^{M} \beta_k \mu_{ik} \nu_{jk}, \tag{6.16}$$

where $M = \min(r - 1, c - 1)$. The parameters satisfy constraints such as

$$\sum_i \mu_{ik} \pi_{i+} = \sum_j \nu_{jk} \pi_{+j} = 0 \qquad \text{for all } k,$$

$$\sum_i \mu_{ik}^2 \pi_{i+} = \sum_j \nu_{jk}^2 \pi_{+j} = 1 \qquad \text{for all } k, \tag{6.17}$$

$$\sum_i \mu_{ik} \mu_{ih} \pi_{i+} = \sum_j \nu_{jk} \nu_{jh} \pi_{+j} = 0 \qquad \text{for all } k \ne h.$$

The model is referred to as the RC(M) *model*.

For $M^* < M$, the special case of model (6.16) with $\beta_k = 0$ for $k > M^*$ is unsaturated. It is called the RC(M^*) *model*. The RC model (6.13) is the case $M^* = 1$. Like the simpler RC model, the RC(M^*) model is also nonlinear, which complicates model fitting, and it treats the variables as nominal unless we replace parameter scores by fixed scores or impose ordering constraints such as in Kateri et al. (1998). Becker (1990a) and Aït-Sidi-Allal et al. (2004) proposed algorithms for ML model fitting. The model is not as simple to interpret as the ordinary RC model and its special cases, so it has not received much attention in applications. Anderson (1996) discussed some ways in which it has been applied.

Goodman (1985, 1996) also noted that the canonical correlation model (6.15) extends to a more general model with sets of parameter scores that is equivalent to the saturated model,

$$\pi_{ij} = \pi_{i+}\pi_{+j}\left(1 + \sum_{k=1}^{M} \rho_k \mu_{ik} \nu_{jk}\right),$$

with constraints for parameters as in the RC(M) model. Here ρ_k is the correlation between μ_{ik} and ν_{jk}, since

$$\sum_i \sum_j \mu_{ik}\nu_{jk}\pi_{ij} = \rho_k.$$

Goodman noted that for this model the Pearsonian distance partitions into

$$\sum_i \sum_j \frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} = \sum_{k=1}^{M} \rho_k^2.$$

Aït-Sidi-Allal et al. (2004) proposed a Fisher scoring algorithm for ML model fitting.

Again, simpler models are possible using $M^* < M$ sets of parameter scores. For example, when ρ_1^2 is nearly as large as $\sum_k \rho_k^2$, it is adequate to use the simpler model (6.15), for which the Pearsonian distance equals ρ_1^2. This is the case for the mental impairment and SES data of Section 6.5.8, for which $M = 3$ and

$$\hat{\rho}_1^2 = 0.0266 = 0.96(\hat{\rho}_1^2 + \hat{\rho}_2^2 + \hat{\rho}_3^2).$$

Based on the Box–Cox transformation, Rom and Sarkar (1992) proposed a generalization of the correlation model and the RC model to

$$\pi_{ij} = \pi_{i+}\pi_{+j}[1 + \phi(\rho u_i v_j)]^{1/\phi}.$$

For this model, X and Y are independent if $\rho = 0$, the ordinary correlation model (6.14) results when $\phi = 1$, and the RC model (6.13) results as ϕ converges to 0 and with parameter scores. The conditional distributions are stochastically ordered, and the correlation for the scores $\{u_i\}$ and $\{v_j\}$ is increasing in ρ for fixed ϕ.

Goodman (1996) proposed a more general model by using an expansion of form $\sum_{k=1}^{M} \beta_k \mu_{ik} \nu_{jk}$ to predict $R(\pi_{ij}/\pi_{i+}\pi_{+j})$ for any monotone-increasing function R defined on the positive real line. Using the log function for R gives association models and using the identity function gives correlation models.

The models of this section also generalize to multiway tables. For example, to describe conditional association in a three-way table, replacing the heterogeneous XY conditional association term $\beta_k u_i v_j$ in model (6.10) by a term $\beta_k \mu_i \nu_j$ with the same parameter scores in each partial table assumes that the form of the association is the same in each partial table but the level of association may vary according to $\{\beta_k\}$ (Xie 1992). The more general term $\beta_k \mu_{ik} \nu_{jk}$ has different parameter scores in each partial table and is equivalent to a separate RC model for each partial table.

6.5.11 Model Selection for Ordinal Variables

In this chapter we have shown several ways to use category orderings in describing associations. With allowance for ordinal effects, the variety of potential models is much greater than standard loglinear models. To choose among models, one approach uses the standard models for guidance. If a standard model fits well, simplify by replacing some parameters with structured terms for ordinal classifications. For example, with three variables, if the standard loglinear model (XY, XZ, YZ) fits well, check whether the fit is still adequate for an ordinal model of homogeneous association such as homogeneous linear-by-linear association.

Association models and correlation models are mainly sensible when we have two ordinal response variables and want to study their association marginally or conditionally on other variables. When one variable alone is a response variable, models presented in earlier chapters are more appropriate.

6.6 MODELING GLOBAL ODDS RATIOS AND OTHER ASSOCIATIONS

For cross-classifications of ordinal variables, the linear-by-linear association model (6.2) can be described by odds ratios for pairs of rows and pairs of columns, as in (6.3). For equally spaced row and column scores, the model implies *uniform association* in terms of the local odds ratios. By contrast, cumulative logit models can be described by odds ratios that are global in the response variable, based on the odds $P(Y \le j)/P(Y > j)$ using cumulative probabilities for each binary collapsing of that variable. With an ordinal explanatory variable with equally spaced scores, the model implies uniform association in terms of cumulative odds ratios. See equation (3.9) in Chapter 3.

6.6.1 Models for Global Odds Ratios

Yet another association model for ordinal variables describes the odds ratios that are global in both variables. Let

$$\theta_{ij}^{G} = \frac{(\sum_{a \le i} \sum_{b \le j} \pi_{ab})(\sum_{a > i} \sum_{b > j} \pi_{ab})}{(\sum_{a \le i} \sum_{b > j} \pi_{ab})(\sum_{a > i} \sum_{b \le j} \pi_{ab})}.$$

A uniform association model assumes constancy of the global odds ratios,

$$\theta_{ij}^{G} = \theta, \qquad i = 1, \ldots, r - 1, \quad j = 1, \ldots, c - 1.$$

The model assumes a common odds ratio for all $(r - 1)(c - 1)$ ways of collapsing the joint distribution into a 2×2 table. Plackett (1965) defined a family of joint distributions that satisfy uniform global odds ratios. The family is constructed from the marginal distributions and the common global odds ratio θ. Given the marginal distribution functions $F_i^X = P(X \leq i)$ and $F_j^Y = P(Y \leq j)$ and a common value $\theta \neq 1$ for $\{\theta_{ij}^{G}\}$, the joint distribution function $F_{ij} = P(X \leq i, Y \leq j)$ is the special case of equation (2.8),

$$F_{ij}$$
$$= \frac{1 + (\theta - 1)(F_i^X + F_j^Y) - \{[1 + (\theta - 1)(F_i^X + F_j^Y)]^2 - 4\theta(\theta - 1)F_i^X F_j^Y\}^{1/2}}{2(\theta - 1)}.$$

When $\theta = 1$, as usual $F_{ij} = F_i^X F_j^Y$, the independence model.

Anscombe (1981, p. 314) discussed ML fitting of the uniform global odds ratio model subject to the constraints that the fitted marginal distributions equal the sample marginal distributions. The sample marginal distributions, together with an estimate for θ, determine an estimate of the joint distribution function using Plackett's construction. Anscombe gave an iterative method for generating a sequence of estimates of θ that converge to the ML estimate. ML fitting is possible with software by treating the model as a member of a class of generalized loglinear models presented in Section 6.6.4, as illustrated for the following example at www.stat.ufl.edu/~aa/ordinal/ord.html.

6.6.2 Example: How Scientific Are Biology and Social Sciences?

Table 6.10 shows responses of those aged 18 to 25 in the 2006 GSS in response to the questions, "How scientific is biology?" and "How scientific is sociology?" The data are sparse, so the sample global odds ratios of $\hat{\theta}_{11}^{G} = 5.11, \hat{\theta}_{12}^{G} = 2.03, \hat{\theta}_{21}^{G} = \infty, \hat{\theta}_{22}^{G} = 2.86$ are unstable estimates. The ML estimate of a uniform global log odds ratio is 0.81 (SE = 0.34), corresponding to a uniform global odds ratio of 2.25 and a 95% confidence interval of (1.2, 4.4).

TABLE 6.10. Results on "How Scientific Are Biology and Sociology?"

Biology	Sociology		
	Very Scientific	Pretty Scientific	Not Scientific
Very scientific	13	80	37
Pretty scientific	1	22	17
Not scientific	0	3	4

Source: 2006 General Social Survey.

The goodness-of-fit statistics are $G^2 = 1.39$ and $X^2 = 1.07$. They are based on df $= 3$, since the model has four odds ratio responses and one parameter. The corresponding fit statistics for the independence model ($\theta^G = 1$) are $G^2 = 7.25$ and $X^2 = 6.47$ (df $= 4$). The likelihood-ratio statistic for testing $\theta^G = 1$ (or $\log \theta^G = 0$) gives a test of independence under the assumption that the uniform global odds ratio model holds. The test statistic equals $7.25 - 1.39 = 5.85$ based on df $= 1$. There is strong evidence of an association (P-value $= 0.02$), which appears to be positive.

6.6.3   Normal Approximations and Model Generalizations

When the uniform global odds ratio model holds, Plackett argued that it is a discrete approximation for a bivariate normal distribution. See Note 6.9. However, Goodman (1981b) showed that a uniform global odds ratio model does not fit as well as a model of uniform local odds ratios when the underlying distribution is bivariate normal and the marginal cutpoints for forming the categories are equally spaced. Global odds ratios do have the advantage, compared to local odds ratios, of being relatively unaffected by the choice of category boundaries (Dale 1984).

More complex models are also possible for the global odds ratios. For example, the model

$$\theta_{ij}^G = \mu_i$$

states that the global odds ratio depends only on the row cutpoint. In collapsing the response, the global odds ratios take into account the ordinal nature of each dimension. Thus, if a model holds for the $\{\theta_{ij}^G\}$, it will generally not hold if rows or columns are permuted. For example, unlike the loglinear row effects model, the model $\theta_{ij}^G = \mu_i$ is not appropriate for nominal row variables.

The modeling approach using global odds ratios shares with the loglinear approach the fact that it treats both variables as response variables. Unlike loglinear models, however, models for global odds ratios do not lend themselves to comparisons of pairs of rows or pairs of columns, because levels are pooled with others rather than being treated individually. Models for global odds ratios are better suited for describing association between two response variables than for comparing pairs of categories of one variable in terms of their conditional distribution on the other variable.

6.6.4   Global Odds Ratio Models as Generalized Loglinear Models

Models for global odds ratios are special cases of a very general family of models. Let $\mathbf{n} = (n_1, \ldots, n_N)'$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$ denote column vectors of observed and expected counts for the N cells of a contingency table. For simplicity we use a single index, but the table may be multidimensional. Loglinear models, including simple association models such as the linear-by-linear association model, have the form

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

for a model matrix $\mathbf{X}$ and a column vector $\boldsymbol{\beta}$ of model parameters. A generalization of this expression provides many additional models. This *generalized loglinear model* is

$$\mathbf{C} \log \mathbf{A} \boldsymbol{\mu} = \mathbf{X} \boldsymbol{\beta} \qquad (6.18)$$

for matrices $\mathbf{C}$ and $\mathbf{A}$. The ordinary loglinear model results when $\mathbf{C}$ and $\mathbf{A}$ are identity matrices.

For global odds ratio models, $\mathbf{C}$ has dimension $(r-1)(c-1) \times 4(r-1)(c-1)$ and $\mathbf{A}$ has dimension $4(r-1)(c-1) \times N$. A particular row of $\mathbf{A}$ has 1 and 0 entries such that, multiplied by $\boldsymbol{\mu}$, it forms a quadrant of expected frequencies used in a particular global odds ratio; there are four such rows for each collapsing of the table to a 2×2 table. A particular row of $\mathbf{C}$ has $+1$ in two places and -1 in two places such as to form the appropriate contrast of the log quadrant counts. For the uniform global log odds ratio model, $\mathbf{X}$ is a $(r-1)(c-1) \times 1$ vector with 1 in each position, and $\boldsymbol{\beta}$ is then the scalar value for the common global log odds ratio.

Other special cases of (6.18) include cumulative logit models, adjacent-categories logit models, and continuation-ratio logit models. For further generalizations, see Lapp et al. (1998) and Coull and Agresti (2003).

ML fitting of generalized loglinear models is not trivial. See Lang and Agresti (1994) and Lang (1996, 2004, 2005). The R function mph.fit developed by Lang can do this (see the Appendix). The fit provides estimated cell probabilities and expected frequencies that satisfy the model.

6.6.5 Models for Measures of Association

When there is a bivariate response, we could formulate models in terms of how a measure of association between the response variables depends on explanatory variables. In a meta-analysis, for example, this could be useful for describing how an effect varies across studies.

Let ζ denote a generic measure of association between two response variables. Let $\zeta(\mathbf{x})$ denote the value of the measure when a vector of explanatory variables takes the value $\mathbf{x}$. A generalized linear model for $\zeta(\mathbf{x})$ has the form

$$\text{link } \zeta(\mathbf{x}) = \alpha + \boldsymbol{\beta}' \mathbf{x}.$$

For instance, the model could specify a uniform value for the global log odds ratio at $\mathbf{x}$. Dale (1986) and Liu (2003) studied a model of this type.

Some authors have combined association models using global odds ratios with regression-type models for each response in terms of explanatory variables. For example, in addition to modeling the association, Dale modeled each response marginal distribution using a cumulative logit model with the same explanatory variables. Molenberghs and Lesaffre (1994), Glonek and McCullagh (1995), Williamson et al. (1995), Glonek (1996), Heagerty and Zeger (1996), Williamson and Kim (1996), and Lapp et al. (1998) also presented models of this type. In Section 9.1.3 we show an example of this generalized type of model.

CHAPTER NOTES

Section 6.1: Ordinary Loglinear Modeling

6.1. Wermuth and Cox (1998) used ordinary loglinear models with ordinal variables by checking whether certain pairs of adjacent categories can be combined. They did this by considering special cases of the models in which certain parameters or contrasts of parameters are set to 0, corresponding to certain types of conditional independence.

Section 6.2: Loglinear Model of Linear-by-Linear Association

6.2. For small samples with the linear-by-linear association model, Agresti et al. (1990) proposed inferences about β. The confidence interval was the method described in Section 2.3.6, and in Section 7.6.1 we present the test of $H_0: \beta = 0$. McDonald et al. (1998) proposed a small-sample goodness-of-fit test. Gross (1981) evaluated the relative efficiencies of various tests of independence in the context of this model. She showed the asymptotic equivalence of the likelihood-ratio test and a correlation-based test of Yates (1948) that uses the same scores, regardless of whether the model truly holds. A test based on rank scores is asymptotically equivalent to these when the model holds with equally spaced scores and the marginal distributions are both uniform.

6.3. If the linear-by-linear association model fits well with small $|\hat{\beta}|$ and if it uses the ridit (average cumulative probability) scores $v_j = (\hat{F}_{j-1}^Y + \hat{F}_j^Y)/2$, the cumulative logit model of proportional odds form with predictor βu_i should have similar fit and have $|\hat{\beta}|$ about half the value of $\hat{\beta}$ for the L × L model. See Note 2.4 for a related remark.

Section 6.3: Row or Column Effects Association Models

6.4. Row effects models were proposed by Haberman (1974), Simon (1974), Duncan and McRae (1979), and Goodman (1979a). Goodman (1979a) proposed an *analysis of association* (ANOAS) as an analog of ANOVA to partition chi-squared statistics pertaining to various aspects of the association. See also Clogg and Shihadeh (1994, pp. 53–61). Chuang et al. (1985), Gilula (1982), and Beh (2001) related association models to the singular value decomposition of a matrix, which itself underlies *correspondence analysis* (Note 6.8).

Section 6.4: Association Models for Multiway Tables

6.5. Clogg and Shihadeh (1994), Ishii-Kuntz (1994), and Etzioni et al. (1994) provided reviews of association models. Articles focusing on such models for multiway tables include Clogg (1982a), Agresti and Kezouh (1983), Goodman (1981c, 1985), Gilula and Haberman (1988), Becker (1989a), Becker and Clogg (1989), Xie (1992), de Falguerolles et al. (1995), and Siciliano and Mooijaart (1997).

6.6. Landis et al. (1978), Koch et al. (1998), and Stokes et al. (2000) reviewed CMH methods. Koch et al. (1982) reviewed related methods. The Mantel–Haenszel

estimate of a common odds ratio for a set of 2×2 tables has been generalized to a set of tables with ordinal response (Liu and Agresti 1996, Liu 2003). An advantage of the generalized CMH methods is that they maintain good performance under sparse asymptotics whereby K grows as n does, possibly with a very small sample size in each stratum. An example is case–control studies with an ordered subclassification of the disease states, in which each stratum provides results for a particular matched set. In such cases, ML model-based estimates can be biased.

Section 6.5: Multiplicative Association and Correlation Models

6.7. The RC model and generalizations of it such as the RC(M) model have been studied by Andersen (1980, pp. 210–216), Becker (1989b, 1990a), Goodman (1979a, 1981a,b, 1985, 1991, 1996), Clogg (1982a,b), Chuang et al. (1985), Chuang and Agresti (1986), Yamaguchi (1990), Simonoff and Tsai (1991), Xie (1992), Anderson (1996), Sobel et al. (1998), Kateri and Iliopoulos (2003), and de Rooij and Heiser (2005). For connections with item response models, see Andersen (1995) and Anderson and Yu (2007). Kateri et al. (2005) and Iliopoulos et al. (2007, 2009) proposed Bayesian inference for the RC model and its generalizations, including order-restricted analyses. Regarding the nonstandard limiting distribution for $G^2(\mathrm{I}) - G^2(\mathrm{RC})$, Hirotsu (1983) gave a related result for analyses making simultaneous comparisons of conditional distributions within rows or within columns of a contingency table. Generalizations of the RC model for multiway tables were presented by Choulakian (1988), Anderson and Böckenholt (2000), Anderson and Vermunt (2000), Becker (1989a), Becker and Clogg (1989), Goodman (1985, 1986, 1996), and Wong (2001). For example, Becker and Clogg (1989) developed such models for comparing associations in several tables. De Rooij (2001) proposed a reparameterization of their models in which association parameters are transformed to distances in multidimensional Euclidean space.

6.8. Kendall and Stuart (1979, Chap. 33) surveyed basic canonical correlation methods for contingency tables. See also Williams (1952), who discussed earlier work by R. A. Fisher and others. The parameter scores in correlation model (6.15) have features in common with scores estimated in the first dimension of a *correspondence analysis*, which is a graphical way of describing association in two-way tables (Goodman 1981a, 1985, 1986, 1996, 2000; Gilula and Haberman 1986; Choulakian 1988; van der Heijden et al. 1989; Gilula and Ritov 1990; Ritov and Gilula 1993; Greenacre 2007). Much research influenced by Goodman's work explored connections among association models, correlation models, and correspondence analysis. See Gilula (1984, 1986); Gilula et al. (1988), Douglas and Fienberg (1990), Ritov and Gilula 1993, Etzioni et al. (1994), Beh (1997), Rayner and Best (2000), and Beh and Davy (2004). Schriever (1983) and Gilula and Ritov (1990) focused on an ordinal version of correspondence analysis pertaining to stochastic orderings of conditional distributions, focusing mainly on models of rank 2 with ordered scores. Beh (1997) and Lombardo et al. (2007) used an alternative approach with orthogonal polynomials.

Section 6.5: Modeling Global Odds Ratios and Other Associations

6.9. Plackett (1965) and Mardia (1967) proposed simple but inefficient ways of fitting the uniform global odds ratio model. Wahrendorf (1980) fitted the model using weighted least squares methods. Plackett stated that the model approximates a bivariate normal distribution with correlation ρ related to the global odds ratio θ by $\rho = cos[\pi/(1 + \sqrt{\theta})]$, for the mathematical constant π. Mardia (1967) studied Plackett's joint distribution and suggested that it better approximates the normal distribution with $\rho = 2 \sin\{\pi(\theta^2 - 1 - 2\theta \log \theta)/[6(\theta - 1)^2]\}$, except close to the means. In central regions, Anscombe (1981, p. 306) suggested the approximation $\rho = (\theta - 1)/(\theta + 1)$.

EXERCISES

6.1. Consider the L × L model (6.2), scaling row scores and column scores to have means of zero and standard deviations of 1. Express the model as a probability function for the cell probabilities $\{\pi_{ij}\}$, and demonstrate the similarity of this function with the bivariate normal density having unit standard deviations and correlation ρ. Show that the association parameter β corresponds to $\rho/(1 - \rho^2)$ (Goodman 1981a,b, 1985).

6.2. For score parameters $\{\mu_i\}$ and $\{\nu_j\}$, explain why the model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta \mu_i \nu_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

is a homogeneous association version of the RC model (6.13). What complications are there in using it to test XY conditional independence?

6.3. An alternative loglinear model for the ordinal–ordinal table, having fixed scores as well as unknown row and column effects, is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i \nu_j + u_i \nu_j.$$

With equally spaced $\{u_i\}$ and $\{v_j\}$, Goodman (1979a, 1981a) referred to it as the R + C *model* because of the additivity in terms of row effects $\{\mu_i\}$ and column effects $\{v_j\}$. Kateri et al. (1998) generalized this model to include both additive and multiplicative effects.

(a) Show that the L × L model, row effects model, and column effects model are special cases.

(b) Specify constraints to make the model identifiable, and show that residual df $= (r - 2)(c - 2)$, like the RC model.

(c) With equally spaced scores, show that the log local odds ratio has the additive form $\log \theta_{ij}^L = \gamma_i + \delta_j$. By contrast, for the RC model (6.13), show that $\log \theta_{ij}^L = \gamma_i \delta_j$.

(d) A multiplicative model for the log global odds ratios has form $\log \theta_{ij}^G = \gamma_i \delta_j$. Is this model invariant to permutations of rows or columns? Why or why not? How does this compare to the RC model?

6.4. Consider a cross-classification of two ordinal variables with fixed row and column scores.

(a) Describe the association pattern represented by the model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta_1 u_i v_j + \beta_2 u_i^2 v_j.$$

For a $3 \times c$ table, explain why this model is equivalent to the row effects model. [More generally, Haberman (1974) expressed the general association term λ_{ij}^{XY} by an expansion using orthogonal polynomials.]

(b) For the model

$$\log \mu_{ij} = \alpha + \lambda_1 u_i + \lambda_2 u_i^2 + \gamma_1 v_j + \gamma_2 v_j^2 + \beta u_i v_j,$$

show that the likelihood equations equate the sample marginal means, marginal standard deviations, and correlation with their fitted values.

6.5. For the homogeneous linear-by-linear association model (6.9), show that the likelihood equation corresponding to β is

$$\sum_k \left(\sum_i \sum_j u_i v_j \hat{\mu}_{ijk} \right) = \sum_k \left(\sum_i \sum_j u_i v_j n_{ijk} \right).$$

Find the likelihood equation corresponding to β_k for the heterogeneous linear-by-linear association model (6.10). What do these equations, together with the equations for the main effects, suggest about observed and fitted correlations?

6.6. Refer to Exercise 2.7. Analyze these data using methods of this chapter.

6.7. Refer to Table 9.1 in Chapter 9. Fit the uniform global odds ratio model for each gender. Show how to compare the associations inferentially.

CHAPTER 7

Non-Model-Based Analysis of Ordinal Association

In this chapter we present ways of describing and making inference about associations between ordinal variables that are not based on models. Here we introduce some measures of association and methods that have been used much longer than the models presented in earlier chapters.

We've seen that $(r-1)(c-1)$ odds ratios in a $r \times c$ contingency table describe the association structure completely. For simplicity, however, it can be useful to summarize the association by a single number. One way to do this is by fitting a model that assumes a common value for all $(r-1)(c-1)$ ordinal odds ratios of a particular type. For example, a model can assume that all the local odds ratios are identical (Section 6.2.2), all the global odds ratios are identical (Section 6.6.1), or all the cumulative odds ratios are identical (Section 3.2.3). An alternative approach, considered in this chapter, uses non-model-based statistics that extend the measures for $2 \times c$ tables introduced in Section 2.1.4.

7.1 CONCORDANCE AND DISCORDANCE MEASURES OF ASSOCIATION

Several measures of association for ordinal variables are based on the numbers of concordant and discordant pairs of observations. From Section 2.2.3, a pair of observations is *concordant* if the subject ranking higher on X also ranks higher on Y. A pair of observations is *discordant* if the subject ranking higher on X ranks lower on Y. Denote the total number of concordant pairs by C and the total number of discordant pairs by D.

7.1.1 Example: Concordant and Discordant Pairs for Happiness Data

We illustrate using the GSS data analyzed in Section 2.2.2 on happiness and family income, shown again in Table 7.1. We treat "very happy" as the high end of the

TABLE 7.1. Happiness and Relative Family Income

	Happiness		
Family Income	Very Happy	Pretty Happy	Not Too Happy
Above average	272	294	49
Average	454	835	131
Below average	185	527	208

scale on happiness. Consider two people, one of whom is classified in the cell (above average, very happy) on (income, happiness) and the other in the cell (average, pretty happy). This pair is concordant, since the first person is ranked higher than the second in both income and happiness. Now each of the 272 persons classified in the cell (above average, very happy) form concordant pairs when matched with each of the 835 people classified (average, pretty happy), so there are $272 \times 835 = 227, 120$ concordant pairs from these two cells. In fact, the 272 people in the cell (above average, very happy) are part of a concordant pair when matched with each of the $(835 + 131 + 527 + 208) = 1701$ people classified lower on both variables.

Figure 7.1 illustrates the pairings of cells that result in concordant pairs. The total number of concordant pairs equals

$$C = 272(835 + 131 + 527 + 208) + 294(131 + 208)$$

$$+ 454(527 + 208) + 835(208) = 1,069,708.$$

$C = 272 (835 + 131 + 527 + 208)$ $\qquad + 294 (131 + 208)$

$+ 454 (527 + 208)$ $\qquad + 835 (208)$
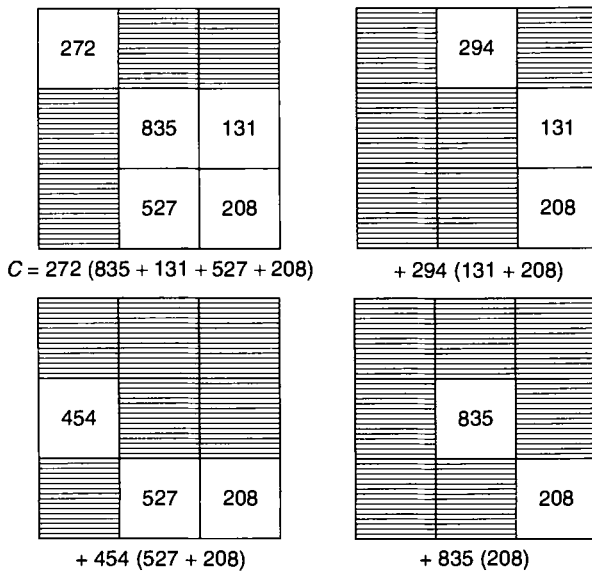
Figure 7.1. Concordant pairs.

By contrast, the total number of discordant pairs equals

$$D = 49(454 + 835 + 185 + 527) + 294(454 + 185)$$
$$+ 131(185 + 527) + 835(185) = 533,662.$$

For these data, $C > D$, which indicates a tendency for those having greater family income to be happier. The formulas for C and D are

$$C = \sum_{i<k}\sum\sum_{j<\ell}\sum n_{ij}n_{k\ell} \quad \text{and} \quad D = \sum_{i<k}\sum\sum_{j>\ell}\sum n_{ij}n_{k\ell}.$$

In each case, the first double summation is over all pairs of rows $i < k$, and the second double summation is over all pairs of columns.

The measures presented next are based on $C - D$, providing various ways of mapping this difference to the interval $[-1, 1]$. For each measure, the association is said to be *positive* if $C - D > 0$ and *negative* if $C - D < 0$. Like the correlation, the measures are most suitable when observations are obtained randomly on both variables. Thus, the population versions are defined in terms of probabilities $\{\pi_{ij}\}$ for a joint distribution.

7.1.2   Gamma

Of the $(C + D)$ pairs of observations that are concordant or discordant, the proportion $C/(C + D)$ is concordant and the proportion $D/(C + D)$ is discordant. The difference between these proportions is called *gamma* (Goodman and Kruskal 1954),

$$\hat{\gamma} = \frac{C - D}{C + D}.$$

For Table 7.1,

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{1,069,708 - 533,662}{1,069,708 + 533,662} = 0.667 - 0.333 = 0.334.$$

Of the pairs that are concordant or discordant, $\frac{2}{3}$ are concordant, $\frac{1}{3}$ are discordant, and the difference between the proportions is 0.334.

The population analog of gamma is

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d},$$

where

$$\Pi_c = 2\sum_{i<k}\sum\sum_{j<\ell}\sum \pi_{ij}\pi_{k\ell} \quad \text{and} \quad \Pi_d = 2\sum_{i<k}\sum\sum_{j>\ell}\sum \pi_{ij}\pi_{k\ell}$$

are the probabilities of concordance and discordance for a randomly selected pair of observations. The factor of 2 occurs in these formulas because the first observation could be in cell (i, j) and the second in cell (k, ℓ), or vice versa.

The value of $\hat{\gamma}$ (and γ) is symmetric, that is, the same whether we treat Y or X or both as response variables. Its range of values is $-1 \le \hat{\gamma} \le 1$, with larger absolute values representing stronger associations. The gamma value just found of 0.334 indicates a relatively weak association. By contrast, when the GSS last asked $X =$ whether you believe in miracles and $Y =$ whether you believe in heaven, using the scale (definitely, probably, probably not, definitely not) for each, $\hat{\gamma} = 0.828$, indicating a strong positive association. The association between belief in heaven and belief in hell was even stronger, with $\hat{\gamma} = 0.895$ and nearly 95% of the untied pairs being concordant.

The boundary values are $\hat{\gamma} = 1$, which occurs when $D = 0$, and $\hat{\gamma} = -1$, which occurs when $C = 0$. Table 7.2 shows contingency tables having various values of $\hat{\gamma}$. The value $|\gamma| = 1$ implies that the relationship is monotone but not strictly monotone. If $\gamma = 1$, in other words, for a pair of observations (X_a, Y_a) and (X_b, Y_b) with $X_a < X_b$, it follows that $Y_a \le Y_b$ but not necessarily that $Y_a < Y_b$. As is true for the correlation, in the population, statistical independence of X and Y implies that $\gamma = 0$, but the converse is not true. For example, γ can equal 0 for a U-shaped bivariate relationship.

For 2×2 tables, $\hat{\gamma}$ simplifies to

$$\hat{\gamma} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

This measure was introduced by the British statistician G. Udny Yule in 1900 and given the symbol Q in honor of Adolphe Quetelet, a Belgian statistician–sociologist–astronomer. Also called *Yule's Q*, it relates to the odds ratio $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$ by

$$\hat{\gamma} = \frac{\hat{\theta} - 1}{\hat{\theta} + 1}.$$

TABLE 7.2. Values of Gamma for Various Cross-Classifications

$\gamma = 1$	$\frac{1}{3}$	0	0
	0	$\frac{1}{3}$	0
	0	0	$\frac{1}{3}$
$\gamma = 1$	0.2	0	0
	0.2	0.2	0
	0	0.2	0.2
$\gamma = 0$	0.2	0	0.2
	0.2	0	0.2
	0	0.2	0
$\gamma = -1$	0.00	0.30	
	0.03	0.67	

For 2×2 tables, gamma is a monotonic function of $\hat{\theta}$ that transforms from a $[0, \infty]$ range onto a $[-1, +1]$ range.

7.1.3 Kendall's Tau-b

Kendall (1945) proposed a measure based on $C - D$ that also uses pairs of observations that are neither concordant nor discordant. These other pairs are *tied* on one or both of the variables, falling in the same row or falling in the same column or both. The number of pairs T_X that are tied on the row variable X and the number of pairs T_Y that are tied on the column variable Y are

$$T_X = \sum_{i=1}^{r} \frac{n_{i+}(n_{i+} - 1)}{2} \quad \text{and} \quad T_Y = \sum_{j=1}^{c} \frac{n_{+j}(n_{+j} - 1)}{2}.$$

The pairs tied on both X and Y are pairs of observations from the same cell. The total number of these pairs is

$$T_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}(n_{ij} - 1)}{2}.$$

For n observations, the total number of pairs decomposes into

$$\frac{n(n - 1)}{2} = C + D + T_X + T_Y - T_{XY}.$$

In this formula, T_{XY} is subtracted because pairs tied on both X and Y are also counted in T_X and in T_Y.

The sample form of *Kendall's tau-b* measure is

$$\hat{\tau}_b = \frac{C - D}{\sqrt{[n(n - 1)/2 - T_X][n(n - 1)/2 - T_Y]}}.$$

The population version is

$$\tau_b = \frac{\Pi_c - \Pi_d}{\sqrt{\left(1 - \sum_i \pi_{i+}^2\right)\left(1 - \sum_j \pi_{+j}^2\right)}}.$$

Like gamma, Kendall's tau-b is symmetric. Since $C + D$ can be no greater than $[n(n - 1)/2 - T_X]$ or $[n(n - 1)/2 - T_Y]$, it also cannot exceed their geometric average, which is the denominator of $\hat{\tau}_b$. Thus, $|\hat{\tau}_b| \le |\hat{\gamma}|$. For 2×2 tables, $\hat{\tau}_b$ is identical to the correlation.

In fact, $\hat{\tau}_b$ is a type of correlation even for $r \times c$ tables, using sign scores for pairs of observations. For each pair of observations (x_a, y_a) and (x_b, y_b), let

$$x_{ab} = \text{sign}(x_a - x_b), \qquad y_{ab} = \text{sign}(y_a - y_b),$$

where $\text{sign}(u) = 1$ if $u > 0$, $\text{sign}(u) = -1$ if $u < 0$, and $\text{sign}(u) = 0$ if $u = 0$. The sign scores $\{x_{ab}\}$ indicate whether x_a is greater than or less than x_b, and similarly for the $\{y_{ab}\}$. Note that $x_{ab} = -x_{ba}$ and $y_{ab} = -y_{ba}$. The product $x_{ab}y_{ab} = 1$ for a concordant pair and -1 for a discordant pair. The square $x_{ab}^2 = 1$ for a pair untied on X and 0 for a pair tied on X. Similarly, $y_{ab}^2 = 1$ for a pair untied on Y and 0 for a pair tied on Y. For the $n(n - 1)$ ordered pairs of observations (a, b) with $a \neq b$,

$$\sum\sum_{a\neq b} x_{ab}y_{ab} = 2(C - D), \qquad \sum\sum_{a\neq b} x_{ab} = \sum\sum_{a\neq b} y_{ab} = 0,$$

$$\sum\sum_{a\neq b} x_{ab}^2 = 2\left[\frac{n(n - 1)}{2} - T_X\right], \qquad \sum\sum_{a\neq b} y_{ab}^2 = 2\left[\frac{n(n - 1)}{2} - T_Y\right].$$

We use each pair twice in these sums, so that $\sum\sum_{a\neq b} x_{ab} = \sum\sum_{a\neq b} y_{ab} = 0$ because of the relationship $x_{ab} = -x_{ba}$ and $y_{ab} = -y_{ba}$. The sample correlation between $\{x_{ab}\}$ and $\{y_{ab}\}$ is therefore

$$\frac{\sum\sum_{a\neq b} x_{ab}y_{ab}}{\sqrt{\left(\sum\sum_{a\neq b} x_{ab}^2\right)\left(\sum\sum_{a\neq b} y_{ab}^2\right)}}$$

$$= \frac{C - D}{\sqrt{\left[n(n - 1)/2 - T_X\right]\left[n(n - 1)/2 - T_Y\right]}} = \hat{\tau}_b.$$

7.1.4   Kendall's Tau

For continuous variables, samples can be fully ranked on both variables, so $T_X = T_Y = T_{XY} = 0$. Then $C + D = n(n - 1)/2$, and $\hat{\gamma}$ and $\hat{\tau}_b$ both simplify to

$$\hat{\tau} = \frac{C - D}{n(n - 1)/2}.$$

In the population, the common value of γ and τ_b is $\Pi_c - \Pi_d$. This measure of ordinal association, proposed by Kendall (1938), is called *Kendall's tau*. Daniels (1944) noted that tau is a correlation coefficient for sign scores, and Kendall (1945) formulated tau-b by constructing the same correlation when ties exist. With categorical data, a nontrivial proportion of the pairs are tied, so tau-b is normally used instead of tau because tau cannot then attain a very large value.

7.1.5   Somers' *d*

Somers (1962) proposed a measure similar to gamma and tau-b, but which treats Y as a response variable and X as an explanatory variable. For this measure, pairs

untied on X serve as the base. The sample version, called *Somers' d*, is

$$d = \frac{C - D}{n(n-1)/2 - T_X}. \tag{7.1}$$

Of the pairs that are untied on X, it is the difference between the proportions of concordant and discordant pairs.

Since the denominator of d is at least as large as the denominator of $\hat{\gamma}$, $|d| \leq |\hat{\gamma}|$. For $|d|$ to equal 1 there must be stricter monotonicity than for $|\hat{\gamma}| = 1$, in the sense that C or D must equal 0, and in addition, none of the pairs that are untied on X can be tied on Y. The population version of Somers' d is

$$\Delta = \frac{\Pi_c - \Pi_d}{1 - \sum_i \pi_{i+}^2}.$$

When $r = 2$, Somers' d is equivalent to the stochastic superiority measure

$$\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$$

for comparing two ordinal distributions introduced in Section 2.1.4. That measure is also useful when the two rows are unordered. For 2×2 tables, Δ simplifies to the difference of proportions, $\pi_{1|1} - \pi_{1|2}$.

7.1.6   Example: Happiness and Income Revisited

For Table 7.1 on $Y =$ happiness and $X =$ family income, in Section 7.1.2 we found that $C = 1,069,708$, $D = 533,662$, and $\hat{\gamma} = 0.334$. Of the $n(n-1)/2 = 2955(2954)/2 = 4,364,535$ pairs in that table, $T_X = 1,619,035$ are tied on family income and $T_Y = 1,859,923$ are tied on happiness. Of the pairs that are untied on family income, the difference between the proportions of concordant and discordant pairs is

$$d = \frac{C - D}{n(n-1)/2 - T_X} = \frac{1,069,708 - 533,662}{4,364,535 - 1,619,035} = 0.195.$$

For pairs of observations, the correlation between the sign scores for family income and the sign scores for happiness is

$$\hat{\tau}_b = \frac{C - D}{\sqrt{[n(n-1)/2 - T_X][n(n-1)/2 - T_Y]}}$$
$$= \frac{1,069,708 - 533,662}{\sqrt{(4,364,535 - 1,619,035)(4,364,535 - 1,859,923)}} = 0.204.$$

All the ordinal measures show a relatively weak tendency for people at higher family income levels to tend to be happier.

7.1.7 Category Choice for Ordinal Variables

Of the measures based on concordant and discordant pairs, gamma is the simplest to interpret. However, in some ways Kendall's tau-b and Somers' d are preferable. As discussed next, gamma tends to be more sensitive than Kendall's tau-b to the choice of the number of categories for the variables and the way they are defined. Also, for $2 \times c$ tables with an ordinal response variable, Somers' d is a useful asymmetric measure.

To illustrate the potential effects of category choice, we collapse the 3×3 table on happiness and family income to a 2×2 table, combining the first two categories of each variable (see Table 2.11). When categories are combined, there are necessarily more ties and fewer concordant and discordant pairs. Gamma takes value 0.50 for the 2×2 table, about 50% larger than its value of 0.33 for the original 3×3 table. Kendall's tau-b is not as greatly affected, changing from 0.20 to 0.19.

Ideally, a measure should be relatively stable with respect to changes in the categorization if it is to be a reliable index of association for cases in which there is no unique way of selecting the categories. As the numbers of rows and columns are increased, there are fewer tied pairs, and gamma and Kendall's tau-b tend to get closer in value to Kendall's tau, which they both equal for continuous variables. We can judge a measure's stability in terms of how close it tends to be to its limiting value for fully ranked data. According to this criterion, gamma fares poorly. It tends to inflate in absolute value as fewer categories are used. When there is an underlying continuous distribution, Kendall's tau-b tends to be closer than gamma to the underlying value of Kendall's tau.

Why does gamma tend to inflate when data are categorized? Pairs of observations that become tied and excluded from calculation in gamma are those that are relatively close on X and/or Y. Ordinarily, pairs of observations selected from a subpopulation in which at least one of the variables is restricted in range exhibit a weaker association than pairs of observations selected randomly from the population. With cruder measurement such pairs are excluded from the calculation of gamma, so the effect is to increase its absolute value (Quade 1974; Agresti 1976).

Not all ordinal methods have biased summaries when crude categorizations are used. For example, when the proportional odds form of the cumulative logit model holds, Section 3.3.3 noted that the parameters describing effects are the same for all ways of categorizing the response variable. Even then, however, cruder categorization tends to result in larger standard errors.

Besides depending on the numbers of categories, the values of most measures of association depend on the marginal distributions of the variables. This is the case for the concordance–discordance measures as well as odds ratios that group categories together. Because of this, it can be risky to compare values of measures calculated in tables having different category definitions or highly different marginal distributions. Consider, for example, case–control studies in which each subject who has a severe case of some disease is matched with someone having a mild case and a set of control subjects not having that disease, with all subjects observed in terms of some exposure that could cause that disease. The expected values of

a summary measure of association would be different for a study that used one control for each pair of cases and a study that used more than one control for each pair of cases. An exception is the local odds ratio. It uses pairs rather than groupings of response categories and maintains the usual invariance to marginal proportions, which is a well-known property of the odds ratio for 2×2 tables.

7.2   CORRELATION MEASURES FOR CONTINGENCY TABLES

In Section 7.1.3 we noted that Kendall's tau-b is a correlation between sign scores for pairs of observations. Other correlation measures are also useful for ordinal data.

7.2.1   Correlation for Fixed or Rank Scores

A common approach treats ordinal variables as of interval-scale type by assigning scores to the rows and columns and using the ordinary correlation. For row scores $u_1 \leq u_2 \leq \cdots \leq u_r$ and column scores $v_1 \leq v_2 \leq \cdots \leq v_c$ and for sample proportions $\{p_{ij}\}$, equation (6.4) showed the sample correlation $\hat{\rho}$. It's simple to find $\hat{\rho}$ using software: Enter the score on each classification for each observation. Then find the ordinary correlation, weighting by the cell count if each observation refers to a cell in a contingency table rather than a single subject.

Alternatively, we could utilize only the ordinal aspect of the variables and use rank-type scores such as ridits (Section 2.1). For sample proportions $\{p_{ij}\}$, the ridit scores for the marginal distributions are the average marginal cumulative proportions,

$$a_i^X = \sum_{k=1}^{i-1} p_{k+} + \frac{p_{i+}}{2}, \qquad i = 1, 2, \ldots, r,$$

$$a_j^Y = \sum_{k=1}^{j-1} p_{+k} + \frac{p_{+j}}{2}, \qquad j = 1, 2, \ldots, c.$$

For each margin, the mean of the ridit scores equals 0.50.

For observations on continuous variables that are ranked on each variable, *Spearman's rho* is the ordinary correlation applied to the rank scores. Kendall (1970, p. 38) proposed an analog of Spearman's rho for contingency tables. It is the ordinary correlation applied using the ridit scores. For sample data it equals

$$\hat{\rho}_b = \frac{\sum_i \sum_j (a_i^X - 0.50)(a_j^Y - 0.50) p_{ij}}{\sqrt{\left[\sum_i (a_i^X - 0.50)^2 p_{i+} \right]\left[\sum_j (a_j^Y - 0.50)^2 p_{+j} \right]}}.$$

Since the ridit scores are a linear function of the midranks, $\hat{\rho}_b$ also equals the correlation applied to the sample midrank scores. For 2×2 tables, $\hat{\rho}_b = \hat{\tau}_b$, and then both measures equal the ordinary correlation.

7.2.2 Contingency Coefficient and Polychoric Correlation

An alternative approach to measuring association assumes a continuous bivariate latent variable underlying the contingency table and approximates the correlation for its distribution. Most of this literature assumes an underlying bivariate normal distribution. Karl Pearson did this in proposing his *tetrachoric correlation* for 2×2 tables, which is the ML estimate based on the four observed frequencies. For $r \times c$ tables, Pearson (1904) proposed an estimate based directly on the chi-squared statistic X^2 for testing independence, his *contingency coefficient*

$$\sqrt{\frac{X^2}{X^2 + n}}.$$

Tallis (1962) proposed finding the ML estimate of the correlation for an assumed underlying bivariate normal distribution for the $r \times c$ table, and the term *polychoric correlation* now usually refers to the ML estimate. The names *polychoric* and *tetrachoric* refer to the names of mathematical expansions Pearson and later authors used in estimating the correlation with categorical data.

Because the underlying random variables are unobserved, the means and variances are arbitrary. For given values, such as means of 0 and standard deviations of 1, the estimable parameters are the correlation and the $r - 1$ row cutpoints and $c - 1$ column cutpoints of the continuous scale that determine the probabilities for the $r \times c$ table. Olsson (1979) proposed an iterative two-step method for finding the ML estimate. In the first step, the cutpoints of the continuous scale that determine the marginal probabilities are estimated using the marginal frequencies. In the second step, the correlation is estimated by solving a likelihood equation for that parameter. Note 7.3 references other articles on this topic.

Given the marginal cutpoints and the correlation estimate, the underlying bivariate normal distribution having that correlation determines joint cell probabilities that have the corresponding marginal probabilities. Multiplying these estimated cell probabilities by n yields fitted values under the assumption of an underlying bivariate normal distribution. These can be compared to the observed cell frequencies with the usual chi-squared statistics to test the goodness of fit of the underlying bivariate normal model, with df $= rc - r - c$. The fit of this model is usually similar to that of the linear-by-linear association model (6.2), which has the marginal frequencies and the correlation (for fixed row and column scores) as sufficient statistics and which tends to fit well when there is underlying bivariate normal distribution (Becker 1989b; Goodman 1981b; Wang 1987, 1997).

7.2.3 Example: Happiness and Income Revisited

Let's reconsider Table 7.1. With equally spaced row and column scores, $\hat{\rho} = 0.223$. The rank-based measure $\hat{\rho}_b = 0.223$ (by coincidence the same, to three decimal places, as $\hat{\rho}$ with equally spaced scores). The ML estimate of the polychoric correlation equals 0.282 (SE $= 0.022$). All these association measures indicate a relatively weak positive association.

TABLE 7.3. Fit of Bivariate Normal Model and Linear-by-Linear Association Model to Table 7.1

Family Income	Happiness		
	Very Happy	Pretty Happy	Not Too Happy
Above average	272 (278.8, 278.7)	294 (299.5, 299.7)	49 (36.7, 36.6)
Average	454 (448.5, 449.1)	835 (807.2, 806.6)	131 (162.7, 164.4)
Below average	185 (183.4, 183.2)	527 (549.6, 549.7)	208 (188.5, 187.1)

Table 7.3. shows the fitted values based on the joint probabilities implied by a bivariate normal distribution with correlation 0.282 that is categorized into a 3×3 table with the given margins. (The fitted values are the first set of parenthesized values.) The deviance chi-squared statistic[1] for testing the model of an underlying bivariate normal distribution equals 14.6, with df = 3. For contrast, Table 7.3 also shows the fit of the linear-by-linear association model with equally spaced row and column scores (the second set of parenthesized values), which has deviance 15.6 with df = 3 and $\hat{\beta} = 0.513$ (SE = 0.043). The fit is similar to that of the bivariate normal model, both showing some lack of fit.

In principle, we could assume a bivariate distribution other than the normal. When the underlying distribution is a bivariate exponential, Edwardes (1993) showed that the underlying values of the correlation and of Kendall's tau for the continuous latent variables are identical.

7.3 NON-MODEL-BASED INFERENCE FOR ORDINAL ASSOCIATION MEASURES

We now present inference methods for ordinal measures of association. We first provide a general result for asymptotic standard errors that are used in confidence intervals. Then we consider significance tests of independence.

A typical sample measure is a function of sample cell proportions in a contingency table. For multinomial sampling, the sample cell proportions have an approximate multivariate normal distribution for large n. The *delta method* implies that the measure itself has a large-sample normal distribution. The variance of that distribution depends on the true cell probabilities and on the partial derivatives of the measure taken with respect to those probabilities. For details about the delta method, see Bishop et al. (1975, Sec. 14.6).

7.3.1 Standard Errors of Ordinal Measures of Association

The ordinal measures of association presented in this chapter take the form of a ratio. The delta method implies the following (Goodman and Kruskal 1972): Let

---

[1]This is available with the *polychor* R function cited in the Appendix.

$\zeta = \nu/\delta$ denote a generic measure of association, with numerator ν and denominator δ that are certain functions of cell probabilities $\{\pi_{ij}\}$. Let

$$\phi_{ij} = \delta\left(\frac{\partial \nu}{\partial \pi_{ij}}\right) - \nu\left(\frac{\partial \delta}{\partial \pi_{ij}}\right).$$

Let $\hat{\zeta}$ denote the sample value of ζ for a multinomial sample. Then as $n \to \infty$, $\sqrt{n}(\hat{\zeta} - \zeta)$ converges in distribution to the normal with mean zero and variance

$$\sigma^2 = \frac{\sum_i \sum_j \pi_{ij}\phi_{ij}^2 - \left(\sum_i \sum_j \pi_{ij}\phi_{ij}\right)^2}{\delta^4}. \tag{7.2}$$

In practice, replacing $\{\pi_{ij}\}$ by their sample values in σ^2 yields the ML estimate $\hat{\sigma}^2$ of σ^2. The term $\mathrm{SE} = \hat{\sigma}/\sqrt{n}$ is an estimated standard error for $\hat{\zeta}$. A Wald confidence interval for ζ is $\hat{\zeta} \pm z_{\alpha/2}(\mathrm{SE})$. As explained in Section 2.3.3, more refined confidence intervals can be based on inverting likelihood-ratio and score tests. See Lang (2008).

The formula for ϕ_{ij} differs from measure to measure, and we defer such formulas to an appendix at the end of this chapter. The standard errors are available with software such as SAS (PROC FREQ), SPSS (CROSSTABS option), Stata (tabulate two-way), and StatXact of Cytel Software (which is also a good source for standard error formulas).

7.3.2   Standard Errors with Independent Multinomial Sampling

Rather than taking a single multinomial sample and cross-classifying it on the variables, some studies take a set of independent multinomial samples. For example, when levels of an explanatory variable refer to groups to be compared, the study might fix the sample size for each group in proportion to its size in the population and then take a random sample of those fixed sizes from the various groups.

Suppose that there is independent multinomial sampling within the rows of a two-way contingency table, with proportion ω_i sampled from row i. When ω_i is the population proportion classified in that row, the sampling is called *proportional sampling*. For a measure of form $\zeta = \nu/\delta$ expressed in terms of conditional probabilities $\{\pi_{j|i}\}$, let

$$\phi_{j|i} = \delta\left(\frac{\partial \nu}{\partial \pi_{j|i}}\right) - \nu\left(\frac{\partial \delta}{\partial \pi_{j|i}}\right).$$

Goodman and Kruskal (1963, 1972) showed that the expression for the asymptotic variance σ^2 of $\sqrt{n}(\hat{\zeta} - \zeta)$ is

$$\sigma^2 = \frac{1}{\delta^4}\sum_i \frac{1}{\omega_i}\left[\sum_j \pi_{j|i}\phi_{j|i}^2 - \left(\sum_j \pi_{j|i}\phi_{j|i}\right)^2\right].$$

Again, $\mathrm{SE} = \hat{\sigma}/\sqrt{n}$ is an estimated standard error.

7.3.3 Testing Independence Using Concordant and Discordant Pairs

We next consider tests of H_0: independence of X and Y. Let $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$ denote the estimated expected frequencies under this hypothesis. The Pearson X^2 and likelihood-ratio G^2 test statistics [see (6.1)] have asymptotic chi-squared null distributions with df $= (r-1)(c-1)$. They are invariant to changes in the order of the rows and/or columns, so they treat both variables as nominal scale and are designed for a general alternative.

In most applications with ordinal variables, associations take the form of a positive or negative trend. That is, Y tends to increase or tends to decrease, in some sense, as X increases. For ordinal measures of association that describe the association by relative numbers of concordant and discordant pairs, positive and negative associations are characterized by the probability orderings $\Pi_c > \Pi_d$ and $\Pi_c < \Pi_d$, respectively. To construct a test having good power for a trend, we can use a test statistic that is natural for H_a: $\Pi_c - \Pi_d > 0$, H_a: $\Pi_c - \Pi_d < 0$, or H_a: $\Pi_c - \Pi_d \neq 0$. Each alternative is narrower than the general one to which the statistics X^2 and G^2 refer, because $\Pi_c \neq \Pi_d$ implies dependence, but the converse is not true.

The quantity $\Pi_c - \Pi_d$ is negative, zero, or positive in precisely the same instances that gamma, Kendall's tau-b, and Somers' d are. Simon (1978) showed that all sample measures having numerator $C - D$ have the same efficacy, and hence the same local power, for testing independence. Thus, it is unnecessary to have separate tests for each measure of association. Instead, a single test can be constructed using $C - D$, the common numerator of their sample values.

For large random samples, $C - D$ is approximately normally distributed. A large-sample test of independence has test statistic

$$z = \frac{C - D}{\sigma_{C-D}}, \tag{7.3}$$

where σ_{C-D} denotes the null standard error of $C - D$. Under H_0: independence, σ_{C-D} depends only on the sample size and the true marginal proportions. Conditional on both sets of sample marginal counts, Kendall (1970, p. 55) showed that

$$\begin{aligned}
\sigma_{C-D}^2 &= \frac{n(n-1)(2n+5) - \sum_i n_{i+}(n_{i+}-1)(2n_{i+}+5)}{18} \\
&\quad - \frac{\sum_j n_{+j}(n_{+j}-1)(2n_{+j}+5)}{18} \\
&\quad + \frac{\left[\sum_i n_{i+}(n_{i+}-1)(n_{i+}-2)\right]\left[\sum_j n_{+j}(n_{+j}-1)(n_{+j}-2)\right]}{9n(n-1)(n-2)} \\
&\quad + \frac{\left[\sum_i n_{i+}(n_{i+}-1)\right]\left[\sum_j n_{+j}(n_{+j}-1)\right]}{2n(n-1)}. \tag{7.4}
\end{aligned}$$

In Exercise 7.3 we summarize the way that this uses the marginal proportions.

Alternatively, we could use a Wald test, such as $z = \hat{\gamma}/\text{SE}$, where SE is the standard error for $\hat{\gamma}$ used for a confidence interval. That is, SE is the *nonnull*

standard error, which is valid whether or not H_0 is true. Rejecting H_0 in favor of a two-sided alternative for such a test at the α level is equivalent to 0 falling outside the corresponding $100(1 - \alpha)\%$ Wald confidence interval. However, under H_0, convergence to the standard normal distribution as n increases is better for the test statistic (7.3) that uses the null standard error, and we recommend that statistic.

7.3.4 Tests Based on Correlation or Pairwise Measures

To detect a positive or negative trend, we could alternatively base tests on a correlation-type measure. Let $\hat{\rho}$ denote a sample correlation, based on assigning fixed scores or rank-type scores to the rows and columns (see Section 7.2). Then, under H_0: independence, the test statistic

$$(n - 1)\hat{\rho}^2 \qquad (7.5)$$

has an asymptotic chi-squared distribution with df $= 1$ (Mantel 1963). In Section 6.4.5 we presented a stratified version of this test.

   Another approach uses a measure that compares each pair of rows simultaneously, using the difference between means or mean ranks. The test statistic uses the sum of such differences for all pairs of rows, maintaining the order of the rows for each comparison. The *Jonckheere–Terpstra test* does this using ranks, combining the results of $r(r - 1)/2$ tests of the Wilcoxon type presented in Section 7.4.1 [see Hollander and Wolfe (1999, pp. 202–210)].

7.3.5 Example: Inference About the Happiness–Income Association

To illustrate inference for ordinal measures, we reconsider the happiness and family income data from Table 7.1. From Section 7.1.2, $C = 1,069,708$, $D = 533,662$, and $\hat{\gamma} = 0.3343$. Formula (7.4) provides $\sigma_{C-D} = 44,022.2$, and hence

$$z = \frac{C - D}{\sigma_{C-D}} = \frac{1,069,708 - 533,662}{44,022.2} = 12.18.$$

There is extremely strong evidence of a positive association.

   Software reports a standard error for gamma of SE $= 0.0262$. A 95% Wald confidence interval for the population value of gamma is

$$0.3343 \pm 1.96(0.0262) \quad \text{or} \quad (0.283, 0.386).$$

This suggests that the association is relatively weakly positive.

   With equally spaced scores for each variable, the sample correlation is $\hat{\rho} = 0.223$. The chi-squared statistic (7.5) for testing independence equals $(n - 1)\hat{\rho}^2 = 146.55$, corresponding to a standard normal statistic of $z = \sqrt{n - 1}\hat{\rho} = 12.11$. Using midrank scores and the Spearman-type correlation $\hat{\rho}_b = 0.223$ for this test yields a nearly identical result.

7.3.6 Sample Size and Power for Establishing an Association

In testing independence with ordinal variables, how large a sample size do we need to obtain a particular power? For comparing $r = 2$ groups, in Section 3.7.2 we presented a formula based on a cumulative logit model of proportional odds form. That formula also applies to the Wilcoxon-type test of Section 7.4.1.

For $r \times c$ tables, most tests are based on asymptotically normal statistics, such as a measure of association (e.g., correlation, gamma) or an estimate of a model parameter. Let $\hat{\zeta}$ denote an asymptotically normal estimator of a generic parameter ζ that equals 0 under the null hypothesis, with variance of the form v/n. Then for a fixed nonnull value ζ_0 of ζ, standard arguments show that the required sample size for a one-sided test with P(type I error) $= \alpha$ and P(type II error) $= \beta$ is

$$n = \frac{(z_\alpha + z_\beta)^2 v}{\zeta_0^2}.$$

To use this formula, the steps are to (1) choose an anticipated set of nonnull cell probabilities, (2) find the value of ζ_0 corresponding to those probabilities, (3) find v for those probabilities, and (4) substitute ζ_0 and v into this formula with the α and β desired. In some cases v has closed form, based on the delta method, and in some cases it requires iterative methods. Even when it has closed form, though, the formula is typically messy computationally. A simple approach to determining v (and ζ_0) enters the anticipated cell probabilities as data into standard software, in which case v equals the square of the reported SE value (since the effective sample size equals 1).

For illustrative purposes, suppose that we had anticipated probabilities proportional to the counts in Table 7.1 relating happiness and family income. In Section 7.3.5 we observed $\hat{\gamma} = 0.3343$ and a standard error of 0.0262 for these data having a sample size of 2955. Setting $v/2955 = (0.0262)^2$ yields $v = 2.028$. To have power 0.90 in a size $\alpha = 0.05$ one-sided test of $\gamma = 0$ when the true relationship has such probabilities with $\gamma_0 = 0.3343$ requires a sample size of about $n = (1.645 + 1.282)^2(2.028)/(0.3343)^2 \approx 156$.

7.3.7 Testing Conditional Independence with Ordinal Variables

To test conditional independence between two ordinal variables for stratified data, ordinal statistics can be constructed that summarize the information about trends in the partial tables. For example, to use concordant and discordant pairs, in partial table k we could find $C_k - D_k$ and its null variance σ_k^2 using equation (7.4). A simple test statistic is then

$$z = \frac{\sum_k (C_k - D_k)}{\sqrt{\sum_k \sigma_k^2}}.$$

In Section 6.4.5 we presented alternative tests for the same purpose using stratum-specific correlation information.

7.4   COMPARING SINGLY ORDERED MULTINOMIALS

When the row variable X is nominal rather than ordinal, the notion of a positive or negative trend is no longer applicable. However, it is still relevant to study whether responses tend to be higher on the ordinal variable Y in some rows than in others. For example, often we expect underlying continuous distributions to be stochastically ordered, with similar variabilities but differing in location.

   We next consider tests of independence that treat the columns as ordinal and the rows as nominal, in the sense that results are invariant to permutations only of the rows. Independence is equivalent to identical population conditional distributions on Y within the r rows. The tests are also useful when the rows are ordinal (rather than nominal) but a positive trend or negative trend is not necessarily expected. Such a test is preferable to the trend tests of Section 7.3 if we expect the levels of X to be stochastically ordered on Y, but not in a monotonic manner, as an example in Section 7.4.4 illustrates.

7.4.1   Comparing Mean Ranks for Two Ordinal Categorical Distributions

We first consider the comparison of two groups ($r = 2$). The best-known nonparametric rank-based test is the *Wilcoxon test* (Lehmann 1975, pp. 18–23). It ranks all the observations on Y and then uses as a criterion the sum of the ranks in the first row relative to its null expectation. For ordered categorical responses, it uses the midranks or ridits.

   Denote the sample sizes by n_1 for row 1 and n_2 for row 2. Denote the sum of the ranks for row 1 by W. Denote the number of tied observations at level j of the response Y by t_j, $j = 1, \ldots, c$. These are the column totals of the $2 \times c$ table. Under the null hypothesis of identical population distributions, conditional on $\{n_i\}$ and $\{t_j\}$,

$$E(W) = \frac{n_1(n_1 + n_2 + 1)}{2},$$

$$\mathrm{Var}(W) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum_{j=1}^{c}(t_j^3 - t_j)}{12(n_1 + n_2)(n_1 + n_2 - 1)}.$$

The second term in the expression for $\mathrm{Var}(W)$ is a correction for ties. It decreases as the number of response categories c and the dispersion of the sample among them increases. (It disappears completely when there are no ties, that is, when all $t_j = 1$.) Under H_0, the test statistic

$$z = \frac{W - E(W)}{\sqrt{\mathrm{Var}(W)}} \tag{7.6}$$

has a large-sample standard normal distribution.

   Equivalent expressions for this test statistic use a numerator that is the difference between the mean ranks or between the mean ridits for the two groups. Such

formulas can use variances expressed in terms of midrank or ridit scores, making it unnecessary to separately correct for ties (Brunner and Puri 2001, Sec. 1.6.1). Here, we use the midrank scores. Let R_{ij} denote the midrank score for observation j in group i, $j = 1, \ldots, n_i$, for $i = 1, 2$. Denote the mean ranks in the two groups by $\bar{R}_1$ and $\bar{R}_2$. The mean rank for the combined sample of $n = (n_1 + n_2)$ observations is $(n + 1)/2$. Then the test statistic equals

$$z = \frac{\bar{R}_1 - \bar{R}_2}{\sqrt{\mathrm{Var}(\bar{R}_1 - \bar{R}_2)}},$$

where

$$\mathrm{Var}(\bar{R}_1 - \bar{R}_2) = \frac{n}{(n-1)n_1 n_2}\left[\sum_{j=1}^{n_1}\left(R_{1j} - \frac{n+1}{2}\right)^2 + \sum_{j=1}^{n_2}\left(R_{2j} - \frac{n+1}{2}\right)^2\right].$$

(7.7)

Alternative expressions compare the mean ridits or compare one of the mean ridits to 0.50 (see Exercise 7.5 and Section 7.4.3 applied to two groups).

The Wilcoxon test is also equivalent to an analysis based on numbers of concordant and discordant pairs. With unordered groups, identifying which are the concordant and which the discordant pairs is arbitrary. This equivalent approach bases inference on the distribution of $C - D$ under the null hypothesis of identical distributions, with the z statistic (7.3). The test using that approach is called the *Mann–Whitney test*. Natural effect measures relating to this test are the α and Δ stochastic superiority measures (Section 2.1.4).

7.4.2 Example: Comparing Treatments for Gastric Ulcer Crater

Table 7.4 shows data from a randomized study to compare two treatments for a gastric ulcer crater. The response was the change in the size of the ulcer crater after three months of treatment, measured with the ordinal scale (larger, less than $\frac{2}{3}$ healed, $\frac{2}{3}$ or more healed, healed). The sample conditional distributions on the ordinal response are (0.19, 0.12, 0.31, 0.38) for treatment A and (0.34, 0.25, 0.25, 0.16) for treatment B.

TABLE 7.4. Results of Study Comparing Two Treatments for Gastric Ulcer

Treatment Group	Change in Size of Ulcer Crater				Total
	Larger	$<\frac{2}{3}$ Healed	$\geq\frac{2}{3}$ Healed	Healed	
A	6	4	10	12	32
B	11	8	8	5	32
Total	17	12	18	17	64

Source: Armitage (1955), with permission of the Biometric Society.

The *cumulative* marginal response counts are (17, 29, 47, 64), so the midranks are

$$\frac{1+17}{2} = 9, \quad \frac{18+29}{2} = 23.5, \quad \frac{30+47}{2} = 38.5, \quad \frac{48+64}{2} = 56.$$

The sum of the ranks for treatment A is

$$W = 6(9) + 4(23.5) + 10(38.5) + 12(56) = 1205.$$

Under the null hypothesis of identical response distributions,

$$E(W) = \frac{32(32+32+1)}{2} = 1040, \quad \sqrt{\text{Var}(W)} = \sqrt{5546.7 - 366.6} = 72.0.$$

The test statistic in terms of the rank sum W is

$$z = \frac{W - E(W)}{\sqrt{\text{Var}(W)}} = \frac{1205 - 1040}{72.0} = 2.29.$$

This has a P-value of 0.02 for the two-sided alternative, giving evidence of a better response with treatment A than with treatment B. In Section 2.1.7 we also analyzed these data and noted that the sample effect was of moderate size, as $\hat{\alpha} = 0.66$ estimates the probability of a better response with treatment A than treatment B. The 95% score confidence interval for α is (0.52, 0.77). The sample size was not very large, so we cannot estimate precisely the size of the effect.

7.4.3 Comparing Mean Ranks for Several Groups

For a continuous response with fully ranked observations, the *Kruskal–Wallis test* is the best known nonparametric rank-based test for comparing r groups (Kruskal and Wallis 1952; Lehmann 1975, pp. 204–210). It is an analysis of variance comparing the r mean ranks. For an ordinal categorical response, a discrete adaptation of the Kruskal–Wallis (KW) test uses midranks. We express the statistic in an equivalent expression using sample mean ridits. The test is designed to detect differences among the r rows in the population mean ridits.

Suppose that the sampling is full multinomial or independent multinomial within the rows. Let $\bar{A}_i$ denote the sample mean ridit for the n_i observations in row i, when the ridits are based on the sample marginal distribution of Y. Then the mean and the variance of $\{\bar{A}_i\}$ are

$$\sum_{i=1}^{r} \frac{n_i}{n} \bar{A}_i = 0.50, \quad \sum_{i=1}^{r} \frac{n_i}{n} (\bar{A}_i - 0.50)^2.$$

The Kruskal–Wallis statistic is proportional to this variance: namely,

$$\text{KW} = \frac{12n}{(n+1)T} \sum_{i=1}^{r} n_i (\bar{A}_i - 0.50)^2, \tag{7.8}$$

where T is a correction factor for ties involving the column totals $\{t_j\}$,

$$T = 1 - \frac{\sum_{j=1}^{c}(t_j^3 - t_j)}{n^3 - n}.$$

Under H_0: identical population distributions in the r rows, the KW test statistic has an asymptotic chi-squared distribution with df $= r - 1$.

For large $\{t_j\}$, T is approximately equal to $1 - \sum_j p_{+j}^3$. It converges upward toward 1.0 as the number of response categories c and the dispersion of the sample among them increases, equalling 1.0 when there are no ties. Alternatively, T can be expressed in terms of the variability in the midrank or ridit scores (Brunner and Puri 2001, Sec. 1.6.1). Let R_{ij} denote the midrank score for observation j in group i, and let $A_{ij} = (R_{ij} - 0.50)/n$ denote the corresponding ridit score. Then, in statistic (7.8),

$$\frac{(n+1)T}{12n} = \frac{1}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(A_{ij} - 0.50)^2 = \frac{1}{n^2(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(R_{ij} - \frac{n+1}{2}\right)^2.$$

7.4.4 Example: Happiness and Number of Sex Partners

Throughout this chapter we've analyzed data on happiness. The GSS in 2006 also asked about the number of sex partners the respondent had in the preceding year. Table 7.5 cross-classifies the number of sex partners (0, 1, 2 or more) with happiness.

We could check for a trend in the data: for example, whether happiness tends to increase as the number of sex partners increases. However, summary ordinal measures of association are near 0, such as $\hat{\gamma} = 0.031$ (SE $= 0.035$) and Kendall's $\hat{\tau}_b = 0.017$ (SE $= 0.020$). A quick look at conditional distributions within rows, also shown in Table 7.5, indicates why. The sample percentage in the "not too happy" category equals 19% for 0 partners, 8% for 1 partner, and 18% for 2 or more partners. Subjects in the first and third rows have a tendency for less happy responses. A monotone trend does not occur.

The Kruskal–Wallis test treats the rows as nominal. It is designed to detect location differences in happiness among the three rows. Software[2] reports

TABLE 7.5. Data on Happiness and Number of Sex Partners

	Happiness		
No. Sex Partners	Not Too Happy	Pretty Happy	Very Happy
0	112 (19%)	329 (55%)	154 (26%)
1	118 (8%)	832 (56%)	535 (36%)
≥2	57 (18%)	198 (63%)	57 (18%)

Source: 2006 General Social Survey.

[2] SAS, using PROC FREQ with the CMH2 and SCORES = RANK options.

KW $= 77.5$, with df $= 2$ (P-value < 0.0001). There is strong evidence that the distribution of happiness is not identical among the three levels for number of sex partners.

7.4.5 Pairwise Comparisons of Groups on an Ordinal Response

In practice, rarely do we believe that the null hypothesis of independence might be true. We expect *some* effects, even if they are small. We learn more from estimating the sizes of any effects: for example, by forming confidence intervals for measures of association.

For the singly ordered analysis, the Kruskal–Wallis test merely indicates whether at least two rows differ in their population mean ridits. Rejection of H_0 suggests two other questions that usually have greater practical importance:

- Which pairs of rows have significant differences? More generally, to what extent can we make inferences about how the rows are ordered on the response?
- How large are the differences among the rows in their response distributions?

To answer the first question, we can repeat the Wilcoxon-type test (7.6) for each pair of rows. This test is equivalent to the KW test applied solely to those two rows, the square of the standard normal statistic (7.6) being the KW statistic (7.8), which then has df $= 1$. Each application of the test uses a different $2 \times c$ table, so the midranks (or ridits) are recomputed each time.

The second question posed above about sizes of differences can be answered by estimating the mean ridit differences or other summary measures for comparing two ordinal categorical distributions, such as

$$\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1), \qquad \alpha = P(Y_1 > Y_2) + \tfrac{1}{2}P(Y_2 = Y_1),$$

introduced in Section 2.1.4. These measures have simple probability interpretations that describe just how different the groups are on the ordinal response. In Section 2.3 we presented confidence intervals for such measures.

7.4.6 Simultaneous Confidence Intervals Comparing Groups

When r is large, there are many pairwise comparisons. A multiple comparison procedure can protect the overall error rate. For the $r(r-1)/2$ pairwise comparisons, suppose that we want the overall type I error rate to be about α. The Bonferroni approach uses the ordinary test but sets the type I error probability for each comparison to be $\alpha/[r(r-1)/2]$; that is, the difference between rows i and k is considered significant if the P-value $\leq 2\alpha/r(r-1)$ for comparing those two rows. This approach is asymptotically a bit conservative, with actual overall error rate bounded above by α.

An alternative method that is less conservative than the Bonferroni method is adapted from a method that Agresti et al. (2008) proposed for comparing several binomial proportions. For the Wilcoxon-type statistic z_{ik} for a pair of rows i and k, declare significance if

$$|z_{ik}| \geq \frac{Q_r(\alpha)}{\sqrt{2}},$$

where $Q_r(\alpha)$ denotes the $100(1 - \alpha)$ percentile of the studentized range distribution with an infinite number of degrees of freedom. (That studentized range distribution is the distribution of the range between the maximum and minimum of r independent standard normal random variables.) In other words, the cutoff point $Q_r(\alpha)/\sqrt{2}$ for the rejection region replaces the value $z_{\alpha/r(r-1)}$ from the standard normal distribution that is used with a Bonferroni adjustment. For $\alpha = 0.05$, for example, the cutoff points for $|z_{ik}|$ when $r = 5$ are 2.81 for the Bonferroni method and 2.73 using the studentized range distribution. With $\alpha = 0.05$, the values of $Q_r(0.05)$ for $r = (2, 3, \ldots, 10)$ are (2.772, 3.314, 3.633, 3.858, 4.030, 4.170, 4.286, 4.387, 4.474).

A similar approach also works with confidence intervals. When a pairwise interval results from inverting a z test that compares an approximately standard normal test statistic to $z_{\alpha/2}$, replace that cutoff point by $Q_r(\alpha)/\sqrt{2}$ to construct a set of $r(r - 1)/2$ confidence intervals that have an overall confidence level approximately equal to $1 - \alpha$. For example, suppose that confidence intervals should have a simultaneous confidence level about 95% for the measure Δ, comparing pairs of groups. A simple way to do this takes the standard errors that software reports for Somers' d (which equals $\hat{\Delta}$ when $r = 2$) and finds a margin of error for Wald-type confidence intervals by multiplying the standard errors by $Q_r(0.05)/\sqrt{2}$. An even better way takes the score-test-based confidence interval alluded to in Section 2.3.3 and replaces $z_{\alpha/2}$ as the critical value for the z form of test statistic by $Q_r(\alpha)/\sqrt{2}$. For further details, see Ryu (2009).

7.4.7   Example: Happiness and Sex Partners Follow-Up

For Table 7.5, the Wilcoxon test for each pair of rows has test statistics $z_{01} = -6.55$ for comparing 0 with one partner, $z_{02} = 1.61$ for comparing 0 with at least two partners, and $z_{12} = 7.25$ for comparing one partner with at least two partners. There is extremely strong evidence that happiness tends to be higher for those with one sex partner than for the other subjects.

The sample estimates of Δ are $\hat{\Delta}_{01} = -0.163$ (SE $= 0.0250$) for comparing 0 with one partner, $\hat{\Delta}_{02} = 0.058$ (SE $= 0.0347$) for comparing 0 with at least two partners, and $\hat{\Delta}_{12} = 0.230$ (SE $= 0.0298$) for comparing one partner with at least two partners. The effects are relatively modest. For approximate simultaneous 95% confidence, we obtain margins of error by multiplying each standard error by $Q_3(0.05)/\sqrt{2} = 2.343$. This yields confidence intervals

$$\Delta_{01} : (-0.22, -0.10), \qquad \Delta_{02} : (-0.02, 0.14), \qquad \Delta_{12} : (0.16, 0.30).$$

7.4.8 Comparing Means with Fixed Scores and Corresponding Trend Tests

Instead of tests comparing mean ranks of multinomial distributions with ordered categories, we could use a test employing fixed scores $\{v_j\}$ for the response categories. One way to do this applies the generalized Cochran–Mantel–Haenszel (CMH) tests of conditional independence presented in Section 6.4.5 to the special case of a single stratum. In this context we use the statistic for unordered rows but ordered columns with scores $\{v_j\}$. The generalized CMH statistic that uses rank column scores corresponds to the square of the Wilcoxon-type statistic (7.6) for two groups and to the Kruskal–Wallis–type statistic (7.8) for several groups.

For the data on comparing treatments for gastric ulcer crater analyzed with the Wilcoxon-type test in Section 7.4.2, the generalized CMH statistic with equally spaced scores for the change in the size of the ulcer crater equals 5.18 with df $= 1$, for a P-value of 0.02. The corresponding generalized CMH statistic with midrank (or ridit) scores equals 5.26, the square root of which is the z statistic of 2.29 for the Wilcoxon-type test reported in Section 7.4.2.

Such tests with fixed scores also relate to binary regression models that reverse the roles of the variables artificially. For the case of two groups, reorient the $2 \times c$ table as a $c \times 2$ table, regarding the groups as the binary response variable Y and the levels of the ordinal variable as the explanatory variable X with scores $\{v_i\}$. Then we treat the data as c independent binomials rather than two independent multinomials, where row i has n_{i1} cases with $y = 1$ out of $n_i = n_{i1} + n_{i2}$ trials. Consider the model

$$\text{link } P(Y = 1 \mid X = i) = \alpha + \beta v_i$$

such as with the logit or probit link. The score test of H_0: $\beta = 0$ is often referred to as the *Cochran–Armitage test* (Armitage 1955; Cochran 1954, 1955). The chi-squared form of the test statistic is

$$\left[\frac{\sum_{i=1}^{c} (v_i - \bar{v}) n_{i1}}{\sqrt{p(1 - p) \sum_{i=1}^{c} n_i (v_i - \bar{v})^2}} \right]^2,$$

where $p = n_{+1}/n$ is the overall proportion of cases with $y = 1$ and $\bar{v} = (\sum_i n_i v_i)/n$ is the sample mean of the fixed scores. The generalized CMH statistic[3] applied to a single stratum with the same scores equals $(n - 1)/n$ times this statistic. Corcoran et al. (2000) compared the power of this test and small-sample conditional tests.

In summary, there are many ways to compare two multinomials with ordered categories, in addition to the model-based methods of Chapters 3 to 5. In Notes 7.5 and 7.6 and the following section we mention yet other methods.

---

[3]The slightly smaller value results from its being derived *conditional* on the number of outcomes of the two types, which is the sufficient statistic for α in the model with logit link (i.e., it is a *conditional score statistic*).

7.5   ORDER-RESTRICTED INFERENCE WITH INEQUALITY CONSTRAINTS

In this chapter inference has utilized ordered categories with summary measures, such as by comparing counts of concordant pairs and discordant pairs or mean ranks. An alternative approach uses order-restricted inference based on inequality constraints for parameters that recognize the ordering. In this section, based on a survey article on this topic by Agresti and Coull (2002), we present ways of doing this using the constraint that ordinal log odds ratios are nonnegative. In Sections 6.3.5 and 6.5.6 in Chapter 6 we presented an alternative order-restricted approach using inequality constraints for parameters in association models.

7.5.1   Inequality Constraints for an Ordinal Predictor of a Binary Response

To begin, consider an $r \times 2$ table with ordered rows. We'll see in Section 7.5.3 that the results also apply to $2 \times c$ tables with ordered columns. Suppose that the rows are independent binomial samples, with "success" counts $(y_1, \ldots, y_r)$ based on $(n_1, \ldots, n_r)$ trials having parameters $(\pi_1, \ldots, \pi_r)$. In many applications, such as in dose–response investigations, we expect that $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_r$ (or $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_r$). Then to test $H_0: \pi_1 = \pi_2 = \cdots = \pi_r$, we can use the order-restricted alternative, $H_a: \pi_1 \leq \pi_2 \leq \cdots \leq \pi_r$.

Denote the sample proportions by $p_i = y_i/n_i$, $i = 1, \ldots, r$. Under H_0, the ML estimator of π_i is the overall sample proportion of successes, $p = \sum_i y_i / \sum_i n_i$. If the r sample proportions satisfy $p_1 \leq p_2 \leq \cdots \leq p_r$, they are the order-restricted ML estimators $\{\hat{\pi}_i\}$ of $\{\pi_i\}$. Otherwise, $\{\hat{\pi}_i\}$ are obtained using the *pooling adjacent violators algorithm*. This pools "out-of-order" pairs of categories for which $p_i > p_{i+1}$ until the resulting sample proportions are monotone increasing. The order-restricted ML estimates $\{\hat{\pi}_i\}$ for the original categories are the sample proportions for the finest partition of categories for which the order restriction occurs.

Denote the hypotheses by I for the null (independence) hypothesis and O for the order-restricted hypothesis. Test statistics compare the fitted counts for I to the fitted counts for O. In row i, the fitted values are $\hat{\mu}_{i1(I)} = n_i p$ and $\hat{\mu}_{i2(I)} = n_i(1 - p)$ under I (i.e., the same proportions in each row) and $\hat{\mu}_{i1(O)} = n_i \hat{\pi}_i$ and $\hat{\mu}_{i2(O)} = n_i(1 - \hat{\pi}_i)$ under O, for the order-restricted ML estimates $\{\hat{\pi}_i\}$. Robertson et al. (1988, p. 167) presented the likelihood-ratio (LR) statistic for testing I against O as a special case of the LR test comparing parameters for independent samples from an exponential family distribution. The test statistic is

$$G^2(\text{I} \mid \text{O}) = 2\sum_{i=1}^{r} n_{i1} \log \frac{\hat{\pi}_i}{p} + 2\sum_{i=1}^{r} n_{i2} \log \frac{1 - \hat{\pi}_i}{1 - p}. \tag{7.9}$$

Equivalently, this can be expressed in the usual LR form for comparing fitted values for two nested multinomial models,

$$G^2(\text{I} \mid \text{O}) = 2\sum_i \sum_j n_{ij} \log \frac{\hat{\mu}_{ij(O)}}{\hat{\mu}_{ij(I)}} = 2\sum_i \sum_j \hat{\mu}_{ij(O)} \log \frac{\hat{\mu}_{ij(O)}}{\hat{\mu}_{ij(I)}}. \tag{7.10}$$

Bartholomew (1959) presented a corresponding Pearson statistic, noting that it equals the ordinary X^2 for testing independence applied to the counts in the collapsed table that combines rows having sample proportions falling out of order.

The large-sample null distribution of $G^2(I \mid O)$ and the corresponding X^2 statistic is *chi-bar-squared*. This is the distribution of a mixture of independent chi-squared random variables of form $\sum_{d=1}^{r} w_d \chi_{d-1}^2$, where χ_d^2 is a chi-squared variate with d degrees of freedom (with $\chi_0^2 \equiv 0$) and w_d is the null probability that the inequality-constrained estimators have d distinct sets on which the estimates are level. For a test statistic T [such as $G^2(I \mid O)$] having this distribution, with observed value t_{obs}, the P-value equals

$$P(T > t_{obs}) = \sum_{d=1}^{r} w_d P(\chi_{d-1}^2 > t_{obs}),$$

the same weighted average of ordinary chi-squared P-values for the possible collapsed tables. Robertson et al. (1988, pp. 74–82) presented $\{w_d\}$ for independent samples from normal populations, and these approximate $\{w_d\}$ for the order-restricted binomial problem. They provided tables of critical values for the chi-bar-squared distribution for $r = 3$ and $r = 4$ (pp. 411–413) and for larger r values for equal-sample-size cases (p. 416). For equal sample sizes with $r = (3, 4, 5, 6, 7, 8)$, and nominal size 0.05, the critical values are (3.82, 4.53, 5.05, 5.46, 5.80, 6.09). Agresti and Coull (1996) presented small-sample tests.

7.5.2 Example: Order-Restricted Treatments in Dose–Response

Table 7.6, from Chuang-Stein and Agresti (1997), refers to a clinical trial for patients who experienced trauma due to subarachnoid hemorrhage. The four ordered treatment groups correspond to a control group and three dose levels of a medication. A study objective was to determine whether a more favorable outcome tends to occur as the dose increases.

To illustrate an inequality-constrained test about binomial parameters, we first compare the four treatment groups on the probability of outcome category 1 (death),

TABLE 7.6. Responses on the Glasgow Outcome Scale from a Clinical Trial with a Placebo (Control) and Three Treatment Groups

Treatment Group	Glasgow Outcome Scale					Total
	Death	Vegetative State	Major Disability	Minor Disability	Good Recovery	
Placebo	59	25	46	48	32	210
Low dose	48	21	44	47	30	190
Medium dose	44	14	54	64	31	207
High dose	43	4	49	58	41	195

Source: Data from Chuang-Stein and Agresti (1997).

combining the other three categories. We test H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4$ against H_a: $\pi_1 \geq \pi_2 \geq \pi_3 \geq \pi_4$. The sample proportions of death were $(p_1, p_2, p_3, p_4) = (0.281, 0.253, 0.213, 0.221)$, so p_3 and p_4 slightly violate the order restriction. The pooling adjacent violators algorithm combines rows 3 and 4 to give $\hat{\pi}_3 = \hat{\pi}_4 = (44 + 43)/(207 + 195) = 0.216$. So the ML fitted probabilities under H_a are $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4) = (0.281, 0.253, 0.216, 0.216)$. The LR statistic for testing H_0 against H_a is $G^2(I \mid O) = 3.27$. From Robertson et al. (1988, pp. 74–82), the $\{\hat{w}_d\}$ estimates are $(0.251, 0.459, 0.249, 0.041)$. The large-sample chi-bar-squared distribution yields a P-value of

$$P = 0.251(0) + 0.459 P(\chi_1^2 \geq 3.27)$$
$$+ 0.249 P(\chi_2^2 \geq 3.27) + 0.041 P(\chi_3^2 \geq 3.27) = 0.095.$$

For comparison, the chi-squared test of independence against the general alternative for the 4×2 table has LR test statistic $G^2(I) = 3.30$ (df $= 3$) and $P = 0.35$. Using the ordering of the dosage levels results in much stronger evidence of an effect.

7.5.3   Inequality-Constrained Odds Ratios in 2×c Contingency Tables

For an $r \times c$ contingency table for which both X and Y are ordinal, in Section 2.2 we introduced various ordinal odds ratios. For any particular ordinal odds ratio, a positive association corresponds to nonnegative values of all $(r - 1)(c - 1)$ log odds ratios. Denote this condition by L for the local odds ratios, G for the global odds ratios, C for the cumulative odds ratios, and Co for the continuation odds ratios. Of these conditions,

L implies C implies G   and   L implies Co implies G.

To conduct order-restricted inference, we could estimate cell probabilities under one of these conditions or test independence against that condition as an alternative.

We discuss this first for comparing two groups ($r = 2$) using a $2 \times c$ table with two independent multinomial samples. The cumulative odds ratios and the global odds ratios are then identical, so conditions C and G are equivalent. They correspond to the conditional distribution of Y being stochastically higher at the second level of X than at the first.

Brunk et al. (1966) derived the ML estimates of the two sets of multinomial probabilities under this stochastic ordering constraint. Let

$$o_j = \frac{\hat{F}_{j|1}}{\hat{F}_{j|2}} = \frac{(n_{11} + \cdots + n_{1j})/n_1}{(n_{21} + \cdots + n_{2j})/n_2}$$

be the ratio of sample cumulative distribution functions at response category j. When the minimum ratio equals 1, which is the value at the final column $j = c$, the sample counts themselves satisfy condition C. Then $\{\hat{\mu}_{ij(O)} = n_{ij}\}$ and the

sample conditional proportions are the order-restricted estimates. Otherwise, the c columns are divided into subsets as follows: The first subset ends at column v_1 for which o_{v_1} is the minimum of $\{o_1, \ldots, o_c\}$. The second subset consists of columns $\{v_1 + 1, \ldots, v_2\}$ such that o_{v_2} is the minimum of $\{o_{v_1+1}, \ldots, o_c\}$, and so on. Grove (1980) provided a geometric representation of this construction. In this construction of subsets, suppose that a particular subset consists of columns $a, a+1, \ldots, b$. Then the ML fitted value under the stochastic ordering restriction for cell (i, j) in those columns equals

$$\hat{\mu}_{ij(O)} = n_{ij} \frac{(n_{+a} + \cdots + n_{+b})n_{i+}}{(n_{ia} + \cdots + n_{ib})n},$$

and at column b, the sample cdf's violate the stochastic order but the fitted cdf's are identical. When a subset contains a single column, $\hat{\mu}_{ij(O)} = (n_{+j}n_{i+})/n$, the fitted value for the independence model.

Grove (1980) and Robertson and Wright (1981) presented the LR test of whether two multinomial distributions are identical against the C alternative. The test statistic has the usual G^2 form (7.10), comparing the fitted values for the null and alternative hypotheses.

The L condition is also called *likelihood-ratio ordering* (Lehmann 1966). Dykstra and Lemke (1988) showed how to obtain the ML fit for this condition. In fact, for $2 \times c$ tables, this fit is identical to the ML fit described in Section 7.5.1 for the alternative of monotone-increasing binomial proportions when we rotate the table and consider the distribution of X given Y. Dykstra et al. (1995) presented the LR test of whether two multinomial distributions are identical against the L alternative. Their test is equivalent to the LR test (7.9) for the rotated table.

The Co alternative has received considerably less attention than L or C. Grove (1984) proposed the LR statistic for comparing two multinomial populations against Co. Oluyede (1993) presented an asymptotically equivalent statistic with Pearson components.

All asymptotic distributions for the various tests for $2 \times c$ tables are chi-bar-squared. The weights $\{w_d\}$ for such distributions are unknown, because they depend on the common unknown multinomial probabilities. Grove (1980) for C and Dykstra et al. (1995) for L suggested that it is adequate to use the approximate P-value that applies when the column totals are equal, for which the end of Section 7.5.1 gave references for critical values.

7.5.4 Inequality-Constrained Odds Ratios in $r \times c$ Contingency Tables

Now we consider $r \times c$ tables with ordered rows and columns. Denote the LR statistics for testing independence against the various order-restricted alternatives for ordinal odds ratios by $G^2(I \mid L)$, $G^2(I \mid C)$, $G^2(I \mid Co)$, $G^2(I \mid G)$. These all have the general form (7.10). Of the conditions $\{L, C, Co, G\}$, the condition L of nonnegative local log odds ratios is the most restrictive, meaning that if L holds, so do the other conditions. The nesting of L within C and Co, which themselves are nested within G, implies that

$$G^2(\text{I} \mid \text{L}) \leq G^2(\text{I} \mid \text{C}) \leq G^2(\text{I} \mid \text{G}) \quad \text{and} \quad G^2(\text{I} \mid \text{L}) \leq G^2(\text{I} \mid \text{Co}) \leq G^2(\text{I} \mid \text{G}).$$

This implies corresponding stochastic orderings of the null distributions. When the sample satisfies L, all four order-restricted fits are identical to the sample data. All four LR test statistics are then identical to $G^2(\text{I})$. Hence, the P-values have the ordering induced by the null distributions, the P-value for $G^2(\text{I} \mid \text{L})$ being smallest. Agresti and Coull (1998) noted that if the true local log odds ratios are strictly positive, the LR test based on them is asymptotically more powerful than the others.

Again, the asymptotic distributions of the LR statistics are chi-bar squared. For the L alternative, Dykstra and Lemke (1988) provided an algorithm for finding the ML fitted values. For the C alternative, no closed form exists for the ML fit when $r > 2$. Wang (1996) showed that the asymptotic chi-bar-squared approximation performs well for small to moderate samples when $\{n_{i+}\}$ are approximately equal but deteriorates as the sample sizes become unbalanced. Wang proposed approximate tests, such as one using the bootstrap. Grove (1984) and Oluyede (1994) proposed large-sample chi-bar-squared tests (not LR) of independence against the Co condition. Dardanoni and Forcina (1998) provided a unified inference for the L and C alternatives. For cell probabilities π, they expressed each alternative in the form $g(\pi) = X\beta$ for $K\beta \geq 0$, where g, X, and K are specific to the type of ordering, and applied a constrained Fisher scoring algorithm to maximize the log-likelihood subject to these constraints. They provided MATLAB programs for large-sample LR testing with L, C, and Co alternatives.

Grove (1986) showed with numerical evaluations that the size of an order-restricted test can be sensitive to the configuration of the marginal probabilities. The choice of test statistic is not obvious unless one is willing to make a rather strong assumption about the nature of the association, and Grove stated that one should select the test statistic designed for the form of association expected. Agresti and Coull (1998) used an optimization program to obtain the ML fits for L, C, Co, and G alternatives and suggested Monte Carlo simulation of exact small-sample conditional tests for the LR test statistics, using the principle presented in Section 7.6.

7.5.5 Example: Dose–Response Revisited

The dose–response data of the 4×5 Table 7.6 has $(4 - 1)(5 - 1) = 12$ ordinal odds ratios of any particular type. All 12 sample global log odds ratios exceed 0, so $G^2(\text{I} \mid \text{G}) = 27.8$ is identical to $G^2(\text{I})$. For the C alternative, two of the 12 sample cumulative odds ratios violate the C order restriction, but barely, and $G^2(\text{I} \mid \text{C}) = 27.7$. For the L alternative, five of the 12 sample local odds ratios violate it, and its fit is somewhat different from the observed data, giving $G^2(\text{I} \mid \text{L}) = 16.1$.

For testing independence with the L, C, or G alternatives, the P-values based on simulated exact conditional LR tests are all at most 0.002. It may seem surprising that this happens for the L alternative, since many sample local odds ratios violate it. However, as mentioned above, its null distribution is stochastically lower, and the

$G^2(I \mid L)$ value is sufficient to provide strong evidence of association ($P = 0.002$). Also, in these tests, a small P-value does not imply that the order restriction truly holds. It implies merely that strong evidence exists against H_0: independence, based on that test criterion.

7.5.6 Anomalies with Order-Restricted Likelihood-Ratio Tests?

Certain ML order-restricted fits and corresponding LR tests can behave in a counterintuitive manner. Even if all the sample log odds ratios are negative, many of the fitted log odds ratios may be strictly positive. A relatively small P-value may occur even though many of the sample odds ratios contradict the order restriction. For example, in Table 7.6 suppose the alternative of interest is G for the reverse ordering of the rows. All sample global log odds ratios are then negative. But 5 of the 12 fitted log odds ratios for the G constraints are then strictly positive, and $G^2(I \mid G) = 11.8$ with $P = 0.26$, even though each sample odds ratio contradicts G.

Cohen and Sackrowitz (1998) claimed that the LR test is inappropriate for various constrained alternatives. Perlman and Wu (1999) argued that counterintuitive results that occur when the sample data fall outside the parameter spaces described by the null and alternative hypotheses may partly reflect a null that is too sharp. They claimed that such results would not occur if testing were, instead, done with a null hypothesis consisting of the entire complement of the order-restricted space rather than a small subset such as independence. Order-restricted inference with sharp null spaces should be used with caution when the sample data seem inconsistent with both hypotheses. Keep in mind that a small P-value does not suggest that the order restriction truly holds but merely that that criterion provides strong evidence against the null.

Of the tests discussed in this section, the one based on $G^2(I \mid L)$ is least likely to provide seemingly anomalous results when the data contradict the order restriction. To illustrate, suppose that every sample local log odds ratio violates L. Then, using properties of related loglinear models, Agresti and Coull (1998) argued that the L fit is identical to the independence fit; thus, $G^2(I \mid L) = 0.0$ and $P = 1.0$, which seems sensible.

7.6 SMALL-SAMPLE ORDINAL TESTS OF INDEPENDENCE

For small samples with $r \times c$ tables, there is a well-known device for conducting tests of independence. The common sampling models are multinomial over the entire table or independent multinomial within the rows or within the columns. Then, conditioning on the row and column marginal totals yields a multivariate hypergeometric null distribution that does not depend on unknown row and/or column marginal probabilities. The probability of a particular table $\{n_{ij}\}$ having the given margins is

$$\frac{(\prod_i n_{i+}!)(\prod_j n_{+j}!)}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}.$$

This is Fisher's method of eliminating nuisance parameters by conditioning on their sufficient statistics.

Methods using this conditional approach are often referred to as *exact* because *P*-values can be calculated without estimating unknown parameters. For 2×2 tables, this method is the basis of *Fisher's exact test*. For $r \times c$ tables and stratified tables, special algorithms and software such as StatXact are available for computing small-sample tests. We recommend these tests when large-sample approximate tests may be invalid.

7.6.1   Small-Sample Tests for Detecting a Trend Association

To construct a test that is sensitive to the category orderings, we order all the tables in the reference set that have the given marginal totals by a statistic T that incorporates the ordering and describes the distance from H_0: independence. For the alternative hypothesis of a positive association, we could take the *P*-value $= P(T \geq t_{obs})$, where T is a correlation or an ordinal measure such as $C - D$ and t_{obs} is its observed value. The *P*-value is the sum of the hypergeometric probabilities for all tables having the given marginal totals for which $T \geq t_{obs}$. For the two-sided alternative, the *P*-value is $P(|T| \geq |t_{obs}|)$. When a statistic T does not have $E(T) = 0$ under H_0, for a two-sided test, order the tables by $|T - E(T)|$ instead.

Agresti and Wackerly (1977), Patefield (1982), and Agresti et al. (1990) proposed such small-sample ordinal tests. Cohen and Sackrowitz (1992) suggested a related test for the one-sided alternative of nonnegative log odds ratios (L), constructing a less discrete *P*-value. Of the tables having $T = t_{obs}$, they included in the *P*-value only those having probability no greater than the probability of the observed table.

7.6.2   Small-Sample Comparisons of Singly Ordered Multinomials

In Section 7.4 we presented large-sample tests for comparing multinomial distributions having ordered categories, using rank scores in discrete versions of Wilcoxon and Kruskal–Wallis test statistics or using fixed scores in trend tests. The exact conditional approach yields small-sample analyses for such statistics. For several multinomials, Klotz and Teng (1977) presented a small-sample analysis with the Kruskal–Wallis statistic (7.8). Mehta et al. (1992) considered a class of tests for comparing two multinomials that includes Wilcoxon-type tests and trend tests with fixed scores. Their analyses also apply in the stratified-data context.

7.6.3   Example: Severity of GVHD in Leukemia Patients

Table 7.7 results from one protocol for a study at the Dana Farber Cancer Institute. For patients receiving a bone marrow transplant, the ordinal response was the severity of graft versus host disease (GVHD). The table, analyzed in Section 2.3.7, showed a suspected risk factor: whether there was a type of blood incompatibility, called a MHC mismatch, between the donor and the recipient of the bone marrow.

The sample sizes for the two groups are small, so we perform a small-sample exact test. The midranks for the five categories are 3, 8.5, 13, 16, 18. The sum

TABLE 7.7. Severity of GVHD in Leukemia Patients by Whether Patient Had MHC Mismatch

MHC Status	Severity of GVHD Toxicity					Total
	None	Mild	Moderate	Severe	Extreme	
Mismatch	2	2	2	1	1	8
Match	3	4	1	2	0	10
Total	5	6	3	3	1	18

Source: StatXact (2005, p. 633), with permission.

of ranks for the mismatch group is 83. StatXact reports that under H_0, given the margin totals this statistic has a distribution with a minimum possible value of 40.5, a maximum possible value of 113.5, a mean value of 76, and a standard deviation of 10.9. The one-sided P-value for the alternative that GVHD severity is worse for the mismatch group is the null probability that the rank sum for the mismatch group is at least 83. This P-value equals 0.28. The two-sided P-value, defined as the null probability of an absolute difference of at least $|83 - 76| = 7$ between the rank sum and its expected value equals 0.57. There is not much evidence of an effect. Unless the true effect is quite large, however, such a small sample size does not provide much power.

7.6.4 Small-Sample Tests of Conditional Independence

Small-sample tests of independence for two-way tables generalize to tests of conditional independence in three-way tables. To eliminate nuisance parameters, we condition on row and column totals in each stratum. This device works for any loglinear model, as these totals are the sufficient statistics for the unknown parameters. The distribution of counts in each stratum is multivariate hypergeometric, and this propagates an exact conditional distribution for the statistic of interest.

For example, for the homogeneous linear-by-linear association model, conditional on the row and column totals in each stratum, the information about the XY association parameter β is contained in its sufficient statistic, which is $\sum_k(\sum_i \sum_j u_i v_j n_{ijk})$. The null hypergeometric distribution in partial table k for the data induces one for $\sum_i \sum_j u_i v_j n_{ijk}$. These then generate a convolution distribution for $\sum_k(\sum_i \sum_j u_i v_j n_{ijk})$, considered over all the partial tables. The P-value for testing H_0: $\beta = 0$ against H_a: $\beta \neq 0$ is the probability of those tables having the same strata margins as observed but test statistic at least as large as observed. See Birch (1965), Agresti et al. (1990), Mehta et al. (1992), and Kim and Agresti (1997).

7.6.5 Dealing with Intense Computations or Discreteness

Exact tests are sometimes impractical when n or the table dimensions are relatively large, because of the enormous number of tables in the reference set having the

same margins as the observed table. In such cases, we can conduct the test using a random sample of tables from the reference set (Agresti et al. 1979). The estimated P-value is then the sample proportion of those tables that have test statistic value at least as large as the observed value. For example, suppose that the actual exact P-value is 0.05, unknown to us, and we estimate it by sampling 1 million tables. Then the sample proportion is an unbiased estimator of the actual P-value. As an estimate, it has standard error $\sqrt{0.05(0.95)/1,000,000} = 0.0002$, small enough for most purposes.

For $r \times c$ tables, Patefield (1982) used Monte Carlo simulation with the LR test statistic for the L alternative, and Agresti and Coull (1998) used it for the L, C, Co, and G alternatives. Agresti and Coull (1996) provided software for simulating exact LR tests comparing two multinomial distributions against the L and C alternatives. Kim and Agresti (1997) used this approach for the generalized CMH test statistics of Section 6.4.5. The software StatXact has the option of simulation to estimate exact P-values precisely.

When n is very small or when the data fall mainly in one row or one column, the conditional distribution of the test statistic T can be highly discrete, having relatively few possible values. Then the exact conditional test can be quite conservative; when H_0 is true, the probability of a P-value ≤ 0.05 may be well below 0.05. To alleviate the conservatism, one can use the *mid P-value*, which is $P(T > t_{\text{obs}}) + \frac{1}{2}P(T = t_{\text{obs}})$. This P-value has null expected value $= 0.50$, like the P-value for a continuous test statistic and unlike the ordinary P-value for discrete data. However, although the test still uses the exact small-sample distribution, the size of the test is no longer guaranteed to be no greater than the nominal value.

7.7   OTHER RANK-BASED STATISTICAL METHODS FOR ORDERED CATEGORIES

Traditional nonparametric methods such as the Wilcoxon test for comparing two groups are formulated for fully ranked data in most texts. They apply to continuous response variables, after replacing the observations by their ranks. Section 7.4 showed that such methods apply to ordered categorical variables using midranks (or ridits). The formulas in this chapter for an ordered categorical response also apply to fully ranked data when we display the data in the form of a contingency table with a separate column for each observation. For example, the Wilcoxon test then applies to a $2 \times n$ table with $n = n_1 + n_2$, each column containing cells with counts of 0 and 1. More generally, Rayner and Best (2001) expressed a variety of nonparametric tests in the context of contingency tables, making it straightforward to handle ordered categorical responses.

In this chapter we have focused on the best known nonparametric methods for bivariate analyses: namely, rank-type correlation analysis when both variables are ordinal and Wilcoxon and Kruskal–Wallis type tests and related summary measures for the singly ordered table. Similarly, nonparametric methods for multivariate analyses can be applied to ordered categorical data.

7.7.1   The Rank Transform Method

Increasingly in recent years, nonparametric rank-based methods have been extended to more complex analyses, such as tests for "no main effects" or for "no interaction" in a two-way layout with independent or matched samples. One strain of this research has used traditional parametric statistics and distributions but with ranks substituted for the observations. Conover and Iman (1981) described this *rank transform* approach.

Later research showed difficulties with this approach (e.g., Akritas 1991), caused by the ranks being nonlinear functions of the data. For example, for the two-way layout, rank statistics are the same for any monotonic transformation of the data, but the truth of hypotheses such as "no interaction" on the scale of the data is not invariant to nonlinear transformations of the data. In addition, homogeneous variances among groups for the original observations is not equivalent to homogeneous variances for the ranks. Akritas and Arnold (1994) showed that rank transform methods are valid when hypotheses are expressed in nonparametric form for effects based on the ranks rather than on the scale of the data. Such effects relate to measures such as mean ridit scores and the stochastic superiority measures α and Δ.

7.7.2   Extended Rank-Based Hypotheses and Inferences

A related methodology has been developed by E. Brunner with various coauthors, building on the Akritas and Arnold research for hypotheses for which relative effects relate to rank-based measures. We mention briefly some of this literature here. For many common settings such as the one-way layout, see Brunner and Puri (2001) for details and Shah and Madden (2004) for a nontechnical summary.

Brunner and Munzel (2000) provided nonparametric solutions for comparing two groups by testing H_0: $\alpha = 0.50$ for the stochastic superiority measure α when the groups are not assumed to have the same variability. Then null is then more general than identical distributions. The asymptotic variance for the test is estimated consistently by using the ranks over all observations as well as the ranks within each sample. An alternative approach is to construct the score test, which is equivalent to checking whether the score confidence interval for α proposed by Ryu and Agresti (2008) contains 0.50.

Munzel and Hothorn (2001) extended the Brunner and Munzel approach for the one-way layout. Brunner and Puri (2001, 2002) presented a general theory for the analysis of nonparametric factorial designs with fixed factors, using hypotheses expressed in terms of distribution functions and test statistics that are score functions.

Motivated by problems arising from multicenter clinical trials, Brunner (1995) considered nonparametric hypotheses and analyses for stratified two-sample designs, including cases where centers and interactions are treated as random effects. The effects are estimated by linear rank statistics using ranks calculated over all centers. Akritas and Brunner (1997) used quadratic forms to test hypotheses about main effects and interactions for a mixed model. Brunner et al. (1997) proposed effective approximations for small-sample distributions of the

quadratic forms used in nonparametric factorial designs. For unbalanced factorial designs, Akritas (1997) proposed tests for nonparametric hypotheses of no main effects, no interaction, and no factor effects. Brunner et al. (1999) developed nonparametric factorial designs for multivariate observations under the framework of general rank-score statistics. They applied the methods to a two-way mixed model assuming compound symmetry and to a factorial design for longitudinal data. For longitudinal data, Brunner and Langer (2000) proposed a nonparametric model that expresses treatment effects and interactions in terms of means of the marginal distributions. The methods proposed extend the Wilcoxon test to factorial designs. For additional methods and examples in the longitudinal setting, see Wei and Lachin (1984) and the text by Brunner et al. (2002).

Bathke and Brunner (2003) proposed a nonparametric alternative to analysis of covariance. Bathke (2005) proposed an asymptotic test for a nonparametric mixed model that is sensitive to detecting a monotone effect of an ordinal covariate. This test contains a test for Spearman's rho as a special case.

Nonparametric methods have traditionally focused mainly on hypothesis testing. Much of the work described above also considers models and estimation of mean ridit-type relative effects and ordering probabilities such as α. Ultimately, this is more informative. Such nonparametric models contrast with the parametric models that are the main focus of this book.

APPENDIX: STANDARD ERRORS FOR ORDINAL MEASURES

This appendix shows the expression for ϕ_{ij} in the variance formula (7.2) based on a multinomial sample, for several ordinal measures of association. In some formulas, it is convenient to use the notation

$$\pi_{ij}^{(c)} = \sum_{a<i}\sum_{b<j} \pi_{ab} + \sum_{a>i}\sum_{b>j} \pi_{ab},$$

$$\pi_{ij}^{(d)} = \sum_{a<i}\sum_{b>j} \pi_{ab} + \sum_{a>i}\sum_{b<j} \pi_{ab}.$$

The term $\pi_{ij}^{(c)}$ is the sum of probabilities for the cells that are concordant when matched with the cell in row i and column j, and $\pi_{ij}^{(d)}$ is the sum of the probabilities for the cells that are discordant when matched with that cell. Figure 7.2 illustrates these probabilities.

Gamma   The population value of gamma is $\gamma = \nu/\delta$ with $\nu = \Pi_c - \Pi_d$ and $\delta = \Pi_c + \Pi_d$. Then

$$\phi_{ij} = 4(\Pi_d \pi_{ij}^{(c)} - \Pi_c \pi_{ij}^{(d)})$$

and $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$, so that $\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2/(\Pi_c + \Pi_d)^4$ [see Goodman and Kruskal (1963, 1972)].
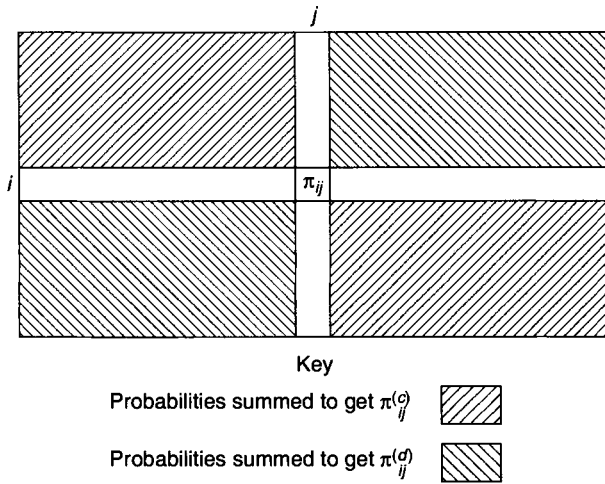
Figure 7.2. Probabilities summed to get $\pi_{ij}^{(c)}$ and $\pi_{ij}^{(d)}$.

Gamma has a tendency to converge slowly to normality and to have distributional irregularity, bias, and skewness problems, especially when the true absolute value is large (Rosenthal 1966; Gans and Robertson 1981). O'Gorman and Woolson (1988) and Carr et al. (1989) showed that better convergence occurs using the Fisher-type transform

$$\hat{\zeta} = \frac{1}{2} \log \frac{1 + \hat{\gamma}}{1 - \hat{\gamma}} \quad \text{for which} \quad \hat{\gamma} = \frac{e^{2\hat{\zeta}} - 1}{e^{2\hat{\zeta}} + 1}.$$

The asymptotic variance of $\hat{\zeta}$ equals the asymptotic variance of $\hat{\gamma}$ multiplied by $(1 - \gamma^2)^{-2}$. A confidence interval can be constructed for ζ and then inverted to one for γ. The measure $\hat{\zeta} = \frac{1}{2} \log(C/D)$ relates to the *ordinal odds ratio* measure C/D proposed by Agresti (1980).

Kendall's Tau-b   The population value of Kendall's tau-b is

$$\tau_b = \frac{\nu}{\delta} \quad \text{with} \quad \nu = \Pi_c - \Pi_d \quad \text{and} \quad \delta = \sqrt{\left(1 - \sum_i \pi_{i+}^2\right)\left(1 - \sum_j \pi_{+j}^2\right)}.$$

For this measure, from Agresti (1976),

$$\phi_{ij} = 2\delta(\pi_{ij}^{(c)} - \pi_{ij}^{(d)}) + \nu\pi_{+j}\sqrt{\left(1 - \sum_a \pi_{a+}^2\right)\left(1 - \sum_b \pi_{+b}^2\right)}$$

$$+ \nu\pi_{i+}\sqrt{\frac{1 - \sum_b \pi_{+b}^2}{1 - \sum_a \pi_{a+}^2}}$$

Somers' d   The population value of Somers' d is

$$\Delta = \frac{\nu}{\delta} \quad \text{with} \quad \nu = \Pi_c - \Pi_d \quad \text{and} \quad \delta = 1 - \sum_i \pi_{i+}^2.$$

For this measure,

$$\phi_{ij} = -2\pi_{i+}(\Pi_c - \Pi_d) - 2\left(1 - \sum_a \pi_{a+}^2\right)(\pi_{ij}^{(c)} - \pi_{ij}^{(d)}).$$

In this case, $\sum_i \sum_j \pi_{ij}\phi_{ij} = -2(\Pi_c - \Pi_d)$ [see Goodman and Kruskal (1972)].

$\Delta = P(Y_2 > Y_1) - P(Y_1 > Y_2)$ For $2 \times c$ tables, Somers' d simplifies to the stochastic superiority measure Δ. It has form $\Delta = \nu/\delta$ with

$$\nu = \sum_{a<b}\sum \pi_{1a}\pi_{2b} - \sum_{a>b}\sum \pi_{1a}\pi_{2b} \quad \text{and} \quad \delta = \pi_{1+}\pi_{2+}.$$

For this measure,

$$\phi_{1j} = \delta\left(\sum_{b>j}\pi_{2b} - \sum_{b<j}\pi_{2b}\right) - \nu\pi_{2+},$$

$$\phi_{2j} = \delta\left(\sum_{a<j}\pi_{1a} - \nu\sum_{a>j}\pi_{1a}\right) - \nu\pi_{1+}, \qquad j = 1, \ldots, c.$$

In this case, $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$. For independent multinomial sampling in the two rows, $\sum_j \pi_{j|1}\phi_{j|1} = \sum_j \pi_{j|2}\phi_{j|2} = 0$. With proportional sampling, the variance is the same as for a single multinomial sample.

$\alpha = P(Y_1 < Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ Since $\alpha = (\Delta + 1)/2$, the standard error of $\hat{\alpha}$ is half the standard error of $\hat{\Delta}$, just considered. Halperin et al. (1989) gave an exact expression for the variance of $\hat{\alpha}$. Equation (2.14) shows its sample estimate. Ryu and Agresti (2008) compared various confidence interval methods for α.

$\theta = \dfrac{P(Y_2 > Y_1)}{P(Y_1 > Y_2)}$ For $2 \times c$ tables, this ordinal odds ratio has form $\theta = \nu/\delta$ with $\nu = \sum\sum_{a<b}\pi_{1a}\pi_{2b}$ and $\delta = \sum\sum_{a>b}\pi_{1a}\pi_{2b}$. For this measure,

$$\phi_{1j} = \delta\sum_{b>j}\pi_{2b} - \nu\sum_{b<j}\pi_{2b} \quad \text{and} \quad \phi_{2j} = \delta\sum_{a<j}\pi_{1a} - \nu\sum_{a>j}\pi_{1a},$$

for which $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$. The measure $\log\hat{\theta}$ converges more rapidly to normality than does $\hat{\theta}$. The asymptotic variance $\tilde{\sigma}^2$ of $\sqrt{n}\log\hat{\theta}$ relates to the asymptotic variance σ^2 of $\sqrt{n}\hat{\theta}$ by $\tilde{\sigma}^2 = \sigma^2/\theta^2$. This result can be used to construct a confidence interval for $\log\theta$. Exponentiating its endpoints yields a confidence interval for θ (Agresti 1980).

CHAPTER NOTES

Section 7.1: Concordance and Discordance Measures of Association

7.1. Kruskal (1958) surveyed many ordinal measures of association, mainly for fully ranked cases. Agresti (1981) extended the stochastic superiority measures to summarize association when comparing several groups. One such measure equals a weighted average of the $|\Delta|$ values for the $2 \times c$ tables for the $\binom{r}{2}$ pairs of groups. Semenya et al. (1983) used a Bradley–Terry type of model (Section 8.6.1) to summarize $\binom{r}{2}$ measures for pairs of groups by $r - 1$ parameters. Agresti et al. (1987b) proposed logit and probit models for the probability that a pair of subjects is concordant. For multiway tables, they modeled the probability that the response at one setting of explanatory variables exceeds the response at another setting. Svensson (2000a,b) defined a measure that compares concordant and discordant pairs for all pairs except those tied on both variables, applying it to describe the association between discrete and continuous scalings of the same variable. Hildebrand et al. (1977) presented ordinal measures of association based on predictive power in using one variable to predict the other. They defined a measure of prediction success as the proportional reduction in prediction error compared to predictions without using the explanatory variable. Measures such as Somers' d and Kruskal's (1958) quadrant measure are special cases.

7.2. Davis (1967), Kendall (1970, p. 121), Quade (1974), and Agresti (1977) used concordant and discordant pairs in measuring ordinal conditional association. Davis proposed $\left[\sum_k (C_k - D_k)\right]/\left[\sum_k (C_k + D_k)\right]$, where C_k and D_k are the numbers of concordant and discordant pairs in partial table k. This is a weighted average of the stratum-specific gamma values. Generalizing Kendall's tau-b as a correlation for sign scores, Somers (1959, 1968), Hawkes (1971), Ploch (1974), and Smith (1974) suggested ordinal analogs of partial slopes and partial correlations by constructing a multiple regression model using sign scores. Beh et al. (2007) used orthogonal polynomials in partitioning measures of association to describe location and dispersion effects for how one or more ordinal variables predicts an ordinal response variable. Carr et al. (1989) applied gamma to longitudinal studies in which an ordinal response variable is observed at least twice for subjects in two or more ordered groups to describe the association between group and response at each time and to evaluate potential group × time interaction on the response.

Section 7.2: Correlation Measures for Contingency Tables

7.3. Mayer and Robinson (1978) presented measures of association between interval-scale and ordinal variables based on maximizing their correlation, considered over all possible monotone transformations of the ordinal variable. To estimate the correlation for an assumed underlying bivariate normal distribution, Ritchie-Scott (1918) proposed a weighted average of tetrachoric correlations for the possible collapsings of the $r \times c$ table to a 2×2 table. Lancaster and Hamdam (1964) showed the inadequacy of Pearson's contingency coefficient estimate of the correlation, and they and Martinson and Hamdan (1972) further investigated the

polychoric correlation. Ronning and Kukuk (1996) summarized Olsson's (1979) polychoric correlation method and considered a more general model. Olsson et al. (1982) extended the polychoric correlation to ML estimation of an underlying correlation when one variable is ordinal and the other is continuous. Drasgow (1988) surveyed polychoric correlations, as did Ekström (2009), who also proposed new ones for larger classes of parametric families. Lee et al. (1992) and Jöreskog (1994) estimated polychoric correlations in the context of structural equation models. This measure is also the basis of multivariate probit models, such as proposed by Kim (1995), Chen and Dey (2000), Biswas and Das (2002), Todem et al. (2007), Webb and Forster (2008), and Lawrence et al. (2008). For other discussions about estimating an underlying normal correlation, see Lancaster (1969, Chap. X), Kendall and Stuart (1979, Chaps. 26 and 9.2), Goodman (1981a,b, 1985), Wang (1987, 1997), and Becker (1989b).

Section 7.3: Non-Model-Based Inference for Ordinal Association Measures

7.4. Yates (1948) suggested a statistic similar to Mantel's correlation statistic (7.5) as well as a statistic that applies with unordered rows (see also Exercise 7.4). Simon (1978) and Gross (1981) analyzed efficiencies for tests based on correlation-type measures. Lipsitz and Fitzmaurice (1996) generalized the correlation test to accommodate missing data. When the data are missing at random but not missing completely at random (Section 9.2.5), they showed that their method is more appropriate than using the ordinary statistic only with the complete cases.

Section 7.4: Comparing Singly Ordered Multinomials

7.5. An alternative nonparametric method for comparing two groups is the Kolmogorov–Smirnov test, based on the maximum absolute difference between the sample cumulative distribution functions (StatXact 2005, pp. 210–217; Hilton et al., 1994; Hilton 1996). For yet other methods, see Gautam (1997), Berger et al. (1998), Cohen et al. (2000), Gautam et al. (2001), and articles cited in Section 7.5. Edwardes (1997) adapted the Wilcoxon test for ordered categories by treating the cutpoints for the outcome categories as randomly determined rather than fixed. This approach relates to random effects models for cumulative probabilities (Chapter 10). Fay and Gennings (1996), Edwardes (2002), and Jung and Kang (2001) extended the Wilcoxon test to handle clustered data.

7.6. For comparing two groups by interchanging variables and using a trend test for a binary response, other tests have been designed that use fixed scores for the ordered categories. Freidlin et al. (1999) and Podgor et al. (1996) used efficiency robustness principles to combine tests from two or more sets of scores into one robust test for analysis in such a way as to minimize the worst possible efficiency loss over all the sets of scores. Zheng (2008) considered the maximum value of a trend statistic over all possible choices of monotone scores and based the P-value on its distribution, thus making the actual decision independent of a choice of scores. In related work, Kimeldorf et al. (1992) used isotonic regression

techniques for binomial responses to find the maximum and minimum values of the sample correlation and of standard test statistics, among the possible values for all possible sets of increasing column scores, and Streitberg and Röhmel (1988) suggested reporting the minimum and maximum P-value found over the possible scorings. Rahlfs and Zimmerman (1993), Ivanova and Berger (2001), and Senn (2007) offered different criteria for choosing scores for a trend test. Ivanova and Berger preferred non-equally-spaced scores to reduce the discreteness of small-sample tests, but Senn argued that equally spaced scores are preferable unless the nature of the categories suggests a different choice.

7.7. Taguchi (1974) proposed a test of independence, called *accumulation analysis*, using cumulative probabilities to compare a set of multinomial distributions. The test statistic is $\sum_{j=1}^{c-1} X_j^2$, where X_j^2 is the Pearson chi-squared statistic for the collapsed response with categories 1 through j in the first column and $j + 1$ through c in the second column. Nair (1986) showed the inadequacies of this approach. Takeuchi and Hirotsu (1982) and Nair (1987) proposed tests that are based on weighted sums $\sum_{j=1}^{c-1} w_j X_j^2$.

Section 7.5: Order-Restricted Inference with Inequality Constraints

7.8. Order-restricted methods have also been proposed for a single set of multinomial probabilities, assuming that $\pi_1 \leq \cdots \leq \pi_c$. Dykstra and Lee (1991) expressed ML estimators in terms of least squares projections. Chacko (1966) tested H_0 : $\pi_1 = \cdots = \pi_c$ against $H_a : \pi_1 \leq \cdots \leq \pi_c$, and Robertson (1978) and Lee (1987) extended Chacko's results to likelihood-ratio tests for and against an arbitrary ordered alternative.

7.9. For other inference relating to stochastic ordering, see Berger (1998), Dardanoni and Forcina (1998), Vermunt (1999), survey articles in a 2002 issue of *J. Statist. Plann. Inference* (Vol. 107, Nos. 1–2), and Chapter 2.1 of Silvapulle and Sen (2004). Agresti and Coull (2002), Gao and Kuriki (2006), and Klingenberg et al. (2009) considered stochastic ordering as an alternative for testing marginal homogeneity with multivariate ordinal data. Jewell and Kalbfleisch (2004) considered a special case of ML estimation with r ordered multinomial distributions in which each of the first $c - 1$ multinomial probabilities have a monotone ordering, applying this to discrete survival times with competing risks. Chuang-Stein and Agresti (1997) surveyed methods for detecting monotone relationships in dose–response relationships. For $2 \times c$ tables, Oh (1995) discussed estimation of cell probabilities under C, L, G, and Co conditions. For testing independence against the G alternative, Nguyen and Sampson (1987) and Rao et al. (1987) presented alternative tests to the likelihood-ratio test. The tests of Nguyen and Sampson are based on the number of tables with the same marginal counts as the table observed that are "more concordant" than the table in the sense of having cdf at least as high in each argument uniformly over all cells in the table. The Rao et al. test uses eigenvalues of a matrix based on dependence ratios $\{p_{ij}/p_{i+}p_{+j}\}$. Kimeldorf and Sampson (1989) provided a unified framework for studying different concepts of positive dependence.

7.10. For the L alternative, Cohen and Sackrowitz (1992) showed that certain ordinal tests, such as one based on numbers of concordant and discordant pairs, are sometimes inadmissible in a decision-theoretic sense. Hirotsu (1982) showed that statistics that have a convex acceptance region and are monotone increasing in each of $s_{ij} = \sum_{a \leq i} \sum_{b \leq j} n_{ab}$ (given the marginal totals) yield efficient score tests. Cohen and Sackrowitz (1991) showed that such tests form a complete class, being the set of exact, unbiased, and admissible tests. A simple way to construct a test in this class is to let the test statistic be some positive linear combination of $\{s_{ij}\}$ and form the critical region from large values of the statistic. For strictly ordered row and column scores, $T = \sum_i \sum_j u_i v_j n_{ij}$ is one such combination. Berger and Sackrowitz (1997) and Cohen and Sackrowitz (1998) discussed a complete class of tests for the C alternative and provided conditions for tests to be unbiased, conditionally unbiased, and in the complete class.

Section 7.6: Small-Sample Ordinal Tests of Independence

7.11. In $2 \times c$ tables, Mehta et al. (1984) proposed small-sample exact tests for nonnull values of local odds ratios as a way of establishing treatment equivalence. For $r \times c$ tables, Bartolucci and Scaccia (2004) used the exact conditional approach with a Pearson statistic, implemented with Markov chain Monte Carlo, to compare the observed counts to the fitted values for an alternative such as L or G. Cohen and Sackrowitz (1991) had considered the L alternative, and Cohen et al. (2003) the C alternative. For surveys of small-sample methods, see the manuals for StatXact (Cytel Software), Agresti (1992b), and Hirji (2005, with pp. 392–394 showing distributions that apply for adjacent-categories logit models with a nominal or quantitative predictor).

EXERCISES

7.1. Show that the *ordinal odds ratio* Π_c / Π_d (Agresti 1980) simplifies for $2 \times c$ tables to $\theta = P(Y_2 > Y_1)/P(Y_1 > Y_2)$, presented in Section 2.1.4, and for 2×2 tables to the odds ratio.

7.2. Let $m = \min(r, c)$ in an $r \times c$ table. Show that the maximum attainable value of $\Pi_c - \Pi_d$ is $(m - 1)/m$, which occurs when the probability is uniformly distributed on m cells in a longest diagonal of the table. Hence,

$$\tau_c = \frac{m(\Pi_c - \Pi_d)}{m - 1}$$

can equal 1.0 in absolute value for any table size (Stuart 1953). Stuart (1963) defined a discrete version of Spearman's rho that can equal 1.0 for any table size: namely,

$$\rho_c = 1 - 6m^2 \sum_i \sum_j \frac{\pi_{ij}(a_i^X - a_j^Y)^2}{m^2 - 1}$$

in terms of the marginal ridit scores. By contrast, Kendall's $|\tau_b|$ and $|\rho_b|$ can equal 1.0 only when $r = c$.

7.3. Refer to formula (7.4) for the null variance of $C - D$. For large marginal totals, show that it equals approximately

$$\tilde{\sigma}_{C-D}^2 = \frac{(1 - \sum_i p_{i+}^3)(1 - \sum_j p_{+j}^3)n^3}{9}.$$

(For Table 7.1, $\tilde{\sigma}_{C-D} = 43,941$ compared to $\sigma_{C-D} = 44,022$.) Show that the estimated null variance of $(\hat{\Pi}_c - \hat{\Pi}_d)$ is approximately

$$\frac{4(1 - \sum_i p_{i+}^3)(1 - \sum_j p_{+j}^3)}{9n},$$

which equals $4/9n$ for Kendall's tau for fully ranked data.

7.4. In an ordinal $r \times c$ table, we could assign ordered scores and use a linear trend model, $E(Y \mid X = u_i) = \alpha + \beta u_i$. With the least squares estimator b of β, show that the Pearson statistic for testing independence partitions into

$$X^2(I) = X^2(L) + X^2(I \mid L),$$

where $X^2(I \mid L) = b^2/\text{Var}(b)$. Here $X^2(L)$ has df $= rc - r - c$ for testing the fit of the linear model, and $X^2(I \mid L)$ has df $= 1$ for testing H_0: independence, given that the linear model holds (Yates 1948). The test based on $X^2(I \mid L)$ is an alternative to tests in Section 7.3. Section 5 presented this model, but using ML to estimate parameters.

7.5. For sample mean ridit scores $\bar{A}_1$ and $\bar{A}_2$ for two groups, show that the Wilcoxon test statistic (7.6) equals

$$z = \frac{\bar{A}_1 - 0.50}{\sqrt{\text{Var}(\bar{A}_1)}} = -\frac{\bar{A}_2 - 0.50}{\sqrt{\text{Var}(\bar{A}_2)}},$$

where

$$\text{Var}(\bar{A}_1) = \frac{n_2(n_1 + n_2 + 1)}{12n_1(n_1 + n_2)^2} - \frac{n_2 \sum_j(t_j^3 - t_j)}{12n_1(n_1 + n_2)^3(n_1 + n_2 - 1)},$$

and $\text{Var}(\bar{A}_2)$ is the formula for $\text{Var}(\bar{A}_1)$ but with n_1 and n_2 interchanged. Moreover,

$$\text{Var}(\bar{A}_1) = \left(\frac{n_2}{n}\right)^2 \text{Var}(\bar{A}_1 - \bar{A}_2),$$

where $\text{Var}(\bar{A}_1 - \bar{A}_2) = (1/n^2)\text{Var}(\bar{R}_1 - \bar{R}_2)$ for $\text{Var}(\bar{R}_1 - \bar{R}_2)$ in equation (7.7). The equivalent expressions for the two mean ridits hold because

$$\frac{n_1}{n}\bar{A}_1 + \frac{n_2}{n}\bar{A}_2 = 0.50 \quad \text{and} \quad n_1^2\text{Var}(\bar{A}_1) = n_2^2\text{Var}(\bar{A}_2).$$

7.6. For drug dosage (low, medium, high) and the effect on a patient's condition (negative, none, positive), the conditional distributions on the effect are expected to exhibit the pattern:

Dosage	Negative	None	Positive
Low	0.30	0.40	0.30
Medium	0.10	0.30	0.60
High	0.30	0.40	0.30

Indicate why it is preferable to test independence using a singly ordered statistic such as the Kruskal–Wallis statistic (7.8) rather than a doubly ordered statistic such as (7.3), even though both variables are ordinal.

7.7. Using the multivariate version of the delta method (Bishop et al. 1975, p. 493), show that the sample mean ridits $\bar{\mathbf{A}} = (\bar{A}_1, \ldots, \bar{A}_{r-1})'$ in a $r \times c$ table have asymptotic covariance matrix $\mathbf{D}'\mathbf{\Sigma}\mathbf{D}/n$, where $\mathbf{\Sigma}/n$ is the $rc \times rc$ multinomial-based covariance matrix for the sample proportions [with $\mathrm{Var}(p_{ij}) = \pi_{ij}(1 - \pi_{ij})/n$ and $\mathrm{Cov}(p_{ij}, p_{k\ell}) = -\pi_{ij}\pi_{k\ell}/n$] and $\mathbf{D}'$ is a $(r - 1) \times rc$ matrix with elements

$$\frac{\partial \bar{A}_k}{\partial \pi_{ij}} = \begin{cases} 1 - a_{j|k}, & i \neq k \\ 1 - a_{j|i} + \dfrac{a_j - \bar{A}_i}{\pi_{i+}}, & i = k, \end{cases}$$

with $a_{j|k} = \sum_{\ell<j} \pi_{\ell|k} + (\pi_{j|k}/2)$. Here $\bar{\mathbf{A}}$ contains only $r - 1$ of the r sample mean ridits because of the linear constraint $\sum_i p_{i+}\bar{A}_i = 0.50$.

7.8. In testing H_0: independence for an $r \times r$ table, suppose that $n_{ii} = 1$ for all i and $n_{ij} = 0$ for all $i \neq j$.

    (a) Show that an exact test that uses the Pearson X^2 to order all tables having the given margins has P-value $= 1.0$.

    (b) Explain why a two-sided test using an ordinal criterion such as $C - D$ or a correlation to order all the tables with the given margins has P-value $= 2/r!$.

7.9. Go to sda.berkeley.edu/GSS and cross-classify the variables POLVIEWS and HAPPY for the latest survey. Using the methods of this chapter, describe the association.

CHAPTER 8

Matched-Pairs Data with Ordered Categories

In this chapter we present methods for matched-pairs data in which each observation in a pair uses the same ordinal scale. The contingency table summarizing the table is then a *square table* for a bivariate response. Matched-pairs data occur when each subject is observed on an ordinal response at two points in time, such as before and after undergoing some experimental treatment or at the beginning and end of a longitudinal study. They also occur when each subject has two sites for observing a response, such as an observation on each eye or on each ear. Sometimes the unit of observation is a couple, such as a social mobility study of parent–child pairs that observes the parent's and the child's social class. Or, each subject might be asked his or her opinion on each of two response variables that have the same scale. Table 8.2, analyzed in several sections of this chapter, is an example of such a square table. It summarizes responses in the 2006 General Social Survey in the United States to the questions "How successful is the government in (1) Providing health care for the sick? (2) Protecting the environment?"

The analysis of matched-pairs responses can have three aspects: comparing the two marginal distributions, analyzing the structure of the joint distribution, and modeling how each response variable (and possibly the association between them) depends on explanatory variables. A comparison of the marginal distributions reveals whether one distribution tends to have higher responses than the other. In Section 8.1 we describe ways of comparing marginal distributions in square tables having ordered categories. In Section 8.2 we show how to conduct the analysis using ordinal models. To analyze the nature of the joint distribution, we could use any of the models discussed in Chapter 6 to investigate the association structure. However, specialized models that have a symmetric structure that recognizes the square nature of the table are usually more appropriate. In Section 8.3 we introduce such models and in Section 8.4 we generalize them to the analysis of matched sets.

In Chapters 9 and 10 we present the two main ways of modeling each response variable in terms of explanatory variables.

In Sections 8.5 and 8.6 we present two applications for which matched-pair models are useful: analyzing agreement between two observers who rate a common set of subjects, and evaluating preferences of treatments based on pairwise comparisons about which is better and by how much.

8.1   COMPARING MARGINAL DISTRIBUTIONS FOR MATCHED PAIRS

Denote the two ordinal responses that are cross-classified as (Y_1, Y_2). Let c denote the number of categories for each response. The data can be summarized in a $c \times c$ contingency table. We regard the cell counts $\{n_{ij} = np_{ij}\}$ with joint sample proportions $\{p_{ij}\}$ as a multinomial sample with sample size n and parameters $\{\pi_{ij}\}$. In this section we present methods for comparing the marginal distributions of the square table.

8.1.1   Marginal Homogeneity in Square Tables

The condition of *marginal homogeneity* states that

$$P(Y_1 = j) = P(Y_2 = j), \qquad j = 1, 2, \ldots, c.$$

For the joint probabilities $\{\pi_{ij}\}$, this says that

$$\pi_{j+} = \pi_{+j}, \qquad j = 1, 2, \ldots, c.$$

Suppose that we want to focus on comparing marginal probabilities for a particular outcome category j. We can test equality of π_{j+} and π_{+j} by collapsing the $c \times c$ table $\{n_{ij}\}$ to a 2×2 table, with categories $(j, \text{not } j)$ in each dimension. For the cell counts $(n_{jj}, \sum_{b \neq j} n_{jb}, \sum_{a \neq j} n_{aj}, \sum_{a \neq j} \sum_{b \neq j} n_{ab})$ in this collapsed table, the test statistic

$$z_j = \frac{\sum_{b \neq j} n_{jb} - \sum_{a \neq j} n_{aj}}{\sqrt{\sum_{b \neq j} n_{jb} + \sum_{a \neq j} n_{aj}}}$$

has an asymptotic standard normal null distribution. This is an application of *McNemar's test*. We could also construct a confidence interval for $(\pi_{j+} - \pi_{+j})$ using the collapsed table. How could we, instead, compare all c pairs of marginal probabilities simultaneously? One way creates a quadratic form by pre- and postmultiplying

$$(p_{1+} - p_{+1}, p_{2+} - p_{+2}, \ldots, p_{(c-1)+} - p_{+(c-1)})$$

by its estimated inverse covariance matrix to obtain a null asymptotic chi-squared variate with df $= c - 1$. However, such an analysis treats the two response variables as nominal.

Ordinal analyses provide more parsimonious description and more powerful inference. This is especially true when c is large, the association between Y_1 and Y_2 is strong, and the focus in on detecting a location shift, that is, whether one response tends to be higher than the other response. We do this using marginal mean scores in this section and using models in Section 8.2. The assessment of sampling variability in such analyses must take into account the matching (rather than independence) of the samples comprising the observed marginal distributions.

8.1.2 Comparing Marginal Mean Scores

To compare marginal mean responses, we assign scores $u_1 \leq u_2 \leq \cdots \leq u_c$ to the outcome categories. Marginal homogeneity implies that $E(Y_1) = E(Y_2)$, where $E(Y_1) = \sum_i u_i \pi_{i+}$ and $E(Y_2) = \sum_i u_i \pi_{+i}$. The sample mean responses are

$$\bar{y}_1 = \sum_i u_i p_{i+} \quad \text{and} \quad \bar{y}_2 = \sum_i u_i p_{+i}.$$

The sample standard error of $\bar{y}_1 - \bar{y}_2$ is

$$\text{SE} = \sqrt{\frac{\sum_i \sum_j (u_i - u_j)^2 p_{ij} - (\bar{y}_1 - \bar{y}_2)^2}{n}}. \tag{8.1}$$

We drop the $(\bar{y}_1 - \bar{y}_2)^2$ term to obtain a null standard error SE_0, as the population value of that term equals 0 under the null hypothesis.

A large-sample confidence interval for $E(Y_1) - E(Y_2)$ is

$$(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}(\text{SE}).$$

A test statistic for testing H_0: marginal homogeneity is

$$z = \frac{\bar{y}_1 - \bar{y}_2}{\text{SE}_0},$$

which has a null asymptotic standard normal distribution. Equivalently, z^2 is a chi-squared statistic with df $= 1$. Meeks and D'Agostino (1983) derived this test as a score test based on a latent variable model. The corresponding statistic replacing SE_0 by SE is a Wald statistic (Bhapkar 1968).

8.1.3 Comparing Marginal Mean Ranks or Mean Ridits

It is also possible to compare the marginal distributions using rank scores. Paralleling the analysis in the preceding section, we could rank the $2n$ observations using midranks and find the mean ranks for the two margins. The midrank for the first category is $[1 + (n_{1+} + n_{+1})]/2$, the average of ranks 1 through $(n_{1+} + n_{+1})$

for the $(n_{1+} + n_{+1})$ observations falling in the first category for either margin. The midrank for category j is

$$r_j = \frac{\left[\left(\sum_{i=1}^{j-1}(n_{i+} + n_{+i})\right) + 1\right] + \sum_{i=1}^{j}(n_{i+} + n_{+i})}{2},$$

which equals $\sum_{i=1}^{j-1}(n_{i+} + n_{+i}) + (n_{j+} + n_{+j} + 1)/2$. The marginal mean ranks are $\overline{R}_1 = \sum_i r_i p_{i+}$ and $\overline{R}_2 = \sum_i r_i p_{+i}$.

The mean ranks do not lend themselves as easily to interpretation as ordinary means for fixed scores, because their size depends on the sample size. Equivalent analyses use ordinal summary measures presented in Section 2.1. Let Y_1 denote an observation from the $\{\pi_{i+}\}$ row marginal distribution and Y_2 an independent observation from the $\{\pi_{+j}\}$ column marginal distribution. The summary measures are the mean *ridits* and corresponding summaries of $P(Y_1 > Y_2)$ and $P(Y_2 > Y_1)$. For the marginal probabilities, Y_1 is stochastically higher than Y_2 if

$$\pi_{1+} + \cdots + \pi_{j+} \leq \pi_{+1} + \cdots + \pi_{+j} \qquad \text{for } j = 1, \ldots, c - 1.$$

A useful measure of the extent to which one marginal distribution is stochastically higher than the other is given by

$$\Delta = P(Y_2 > Y_1) - P(Y_1 > Y_2).$$

This measure is positive when Y_2 is stochastically higher than Y_1 and negative when Y_1 is stochastically higher than Y_2. An alternative, equivalent measure is

$$\alpha = P(Y_2 > Y_1) + \frac{1}{2}P(Y_1 = Y_2) = \frac{\Delta + 1}{2}.$$

Marginal homogeneity implies that $\Delta = 0$ and $\alpha = \frac{1}{2}$.

In Sections 2.3 and 7.4.1 we presented inference for the stochastic superiority measures Δ and α for two independent multinomial samples in a $2 \times c$ table. Now, by contrast, we estimate them for a $2 \times c$ table consisting of the marginal distributions of a $c \times c$ table, so that the samples in the two rows are matched rather than independent. In terms of the $c \times c$ table, the sample value of Δ is

$$\hat{\Delta} = \sum_{i<j}\sum p_{i+}p_{+j} - \sum_{i>j}\sum p_{i+}p_{+j}.$$

See Agresti (1983b), Svennson (1997, 1998), and Svennson and Holm (1994).

The sample versions of these measures relate to ridit scores. The ridit score in category j using the sample marginal distribution of Y_1 is

$$a_{1j} = \hat{P}(Y_1 < j) + \frac{1}{2}\hat{P}(Y_1 = j) = p_{1+} + \cdots + p_{(j-1)+} + \frac{1}{2}p_{j+}.$$

The ridit score in category j using the sample marginal distribution of Y_2 is

$$a_{2j} = p_{+1} + \cdots + p_{+(j-1)} + \tfrac{1}{2}p_{+j}.$$

Let $\bar{A}_U(V)$ denote the mean ridit for the distribution of V when we use the distribution of U to define the ridits. So for Y_1 and Y_2,

$$\bar{A}_{Y_1}(Y_2) = \sum_j p_{+j}a_{1j} = \sum_j p_{+j}\left(p_{1+} + \cdots + p_{(j-1)+} + \frac{p_{j+}}{2}\right),$$

$$\bar{A}_{Y_2}(Y_1) = \sum_j p_{j+}a_{2j} = \sum_j p_{j+}\left(p_{+1} + \cdots + p_{+(j-1)} + \frac{p_{+j}}{2}\right).$$

Then the ordinal measures $\hat{\Delta}$ and $\hat{\alpha}$ relate to these mean ridits by (Vigderhous 1979, Agresti 1983b)

$$\Delta = \bar{A}_{Y_1}(Y_2) - \bar{A}_{Y_2}(Y_1) \quad \text{and} \quad \hat{\alpha} = \bar{A}_{Y_1}(Y_2).$$

8.1.4 Inference about Δ and α for Matched Pairs

The estimated large-sample variance of $\hat{\Delta}$ for matched multinomial distributions induced by a multinomial distribution over the $c \times c$ table is

$$\mathrm{SE}_{\hat{\Delta}}^2 = \frac{\sum_i \sum_j \hat{\phi}_{ij}^2 p_{ij} - \left(\sum_i \sum_j \hat{\phi}_{ij} p_{ij}\right)^2}{n},$$

where $\hat{\phi}_{ij} = 2(a_{1j} - a_{2i})$. The standard error of $\hat{\alpha}$ is $\mathrm{SE}_{\hat{\alpha}} = (\mathrm{SE}_{\hat{\Delta}})/2$. For large n, the null hypothesis of marginal homogeneity can be tested using the statistic

$$z = \frac{\hat{\Delta}}{\mathrm{SE}_{\hat{\Delta}}} = \frac{\hat{\alpha} - 0.50}{\mathrm{SE}_{\hat{\alpha}}},$$

which has approximately a standard normal null distribution. When the true marginal distributions are stochastically ordered, this test and the test comparing means can be much more powerful than a chi-squared test having df $= c - 1$ for comparing the margins. The power advantage increases as c increases (Agresti 1983b).

Using the SE values, we can construct Wald confidence intervals for Δ or α. When the true effects are strong, so that these measures fall near their boundary values, the actual coverage probabilities for such intervals need not fall near their nominal levels. Ryu and Agresti (2008) noted that a better Wald approach constructs the interval for logit (α) and then inverts it to the α or Δ scale. The Wald confidence interval for logit (α) is

$$\text{logit } (\hat{\alpha}) \pm \frac{z_{\alpha/2}(\mathrm{SE}_{\hat{\alpha}})}{\hat{\alpha}(1 - \hat{\alpha})}.$$

Its bounds [LB, UB] induce the interval

$$
\left[\frac{\exp(\text{LB})}{1 + \exp(\text{LB})}, \frac{\exp(\text{UB})}{1 + \exp(\text{UB})} \right]
$$

for α. Simulations suggest that this method performs well, although more sophisticated methods such as a score-test-based or profile likelihood confidence interval are needed when the sample values fall at the boundary.

8.1.5 Example: Occupational Mobility in Britain

Table 8.1 relates fathers' and sons' occupational status category for a British sample. Except for the (n_{13}, n_{31}) combination, the sample satisfies $n_{ij} < n_{ji}$ whenever $i < j$. Thus, for each other pair of categories, the proportion of the father–son pairs for which the son had the higher status category is greater than the proportion for which the father had the higher status category. The sample marginal distributions are stochastically ordered, and there is a marked tendency for more sons to have occupations with the highest status. For the population represented by this sample, we analyze whether the occupational distribution for sons differs from the occupational distribution for fathers.

With scores equal to the status numbers, the marginal sample means are $\bar{y}_1 = 3.701$ for fathers and $\bar{y}_2 = 3.799$ for sons. The difference has $\text{SE} = 0.01972$ and $\text{SE}_0 = 0.01974$. Since $z = (\bar{y}_1 - \bar{y}_2)/\text{SE}_0 = 4.5$, there is strong evidence of a difference between the population marginal means. The large sample size results in the narrow confidence interval (0.05, 0.13) for the difference between the population means for sons and fathers.

For the ordinal rank-based approach, $\hat{\Delta} = 0.0507$ and $\hat{\alpha} = 0.525$. For H_0: $\Delta = 0$ (and $\alpha = \frac{1}{2}$), the test statistic $z = \hat{\Delta}/\text{SE}_{\hat{\Delta}} = 0.0507/0.0102 = 5.0$ also gives extremely strong evidence of marginal inhomogeneity. A 95% confidence interval for Δ is (0.031, 0.071). The corresponding confidence interval for α is $[(0.031+1)/2, (0.071+1)/2)] = (0.505, 0.545)$. We conclude that there is a slight tendency for a son's occupational status to be higher than a father's.

TABLE 8.1. Occupational Status for British Father–Son Pairs

Father's Status[a]	Son's Status[a]					Total
	1	2	3	4	5	
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1510
5	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

Source: D. V. Glass, ed., *Social Mobility in Britain*, Free Press, New York, 1954, with permission.
[a] 1, lowest status; 5, highest status.

8.2 MODELS COMPARING MATCHED MARGINAL DISTRIBUTIONS

In Section 8.1 we compared the marginal distributions directly. A test comparing marginal means is a test of H_0: $\beta = 0$ for the simple model

$$E(Y_1) = \alpha, \qquad E(Y_2) = \alpha + \beta.$$

Alternative models directly address the categorical nature of the response scale. Any type of model introduced in previous chapters is valid as long as the analysis accounts for the dependence between the two sample marginal distributions. We illustrate by formulating models for cumulative logits.

8.2.1 Cumulative Logit Model Comparing Marginal Responses

A cumulative logit model for two marginal distributions that may differ in their location is

$$\text{logit}\,[P(Y_1 \leq j)] = \alpha_j \quad \text{and} \quad \text{logit}\,[P(Y_2 \leq j)] = \alpha_j + \beta \qquad (8.2)$$

for $j = 1, \ldots, c - 1$. The model makes the proportional odds assumption by which the effect β is the same for each cumulative probability. For each j, the odds of outcome $Y_2 \leq j$ equal $\exp(\beta)$ times the odds of outcome $Y_1 \leq j$. The model implies stochastically ordered marginal distributions, with $\beta > 0$ when Y_1 tends to be higher than Y_2. Marginal homogeneity corresponds to $\beta = 0$. As usual, $\{\alpha_j\}$ are monotone increasing in j since the cumulative probabilities increase in j.

Model fitting treats (Y_1, Y_2) as dependent. The ML approach maximizes the multinomial likelihood function for the joint distribution of cell probabilities $\{\pi_{ij}\}$ in the $c \times c$ table. This is not simple to do with the standard model-fitting functions in ordinary software. The model refers to marginal probabilities $\{P(Y_1 = i) = \pi_{i+}\}$ and $\{P(Y_2 = j) = \pi_{+j}\}$, but the multinomial log likelihood, $\sum_i \sum_j n_{ij} \log \pi_{ij}$, contains joint probabilities. To illustrate, the example in Section 8.2.2 fits this model to the two margins of a 3×3 contingency table. The natural sampling model is a multinomial distribution over the nine cells of the table, which is a joint distribution for the two responses. Since the model refers to the marginal probabilities, it is not possible to substitute the model formula into the likelihood function as is done with univariate response models to generate a function of the model parameters that the ML estimates maximize.

We defer discussion of ML model fitting of marginal models to Section 9.1.2. Specialized software is available, such as the mph.fit R function described in the Appendix that maximizes the multinomial likelihood function by treating the model formula as a set of constraint equations for the maximization. In Section 9.2 we present a simpler (not ML) way to estimate model parameters that is available in ordinary software: the generalized estimating equations (GEE) approach.

Model (8.2) describes the $2(c - 1)$ marginal probabilities by c parameters, so df $= c - 2$ for testing fit. A test of marginal homogeneity has H_0: $\beta = 0$. A summary of the degree of marginal inhomogeneity is given by a confidence interval for β or for the cumulative odds ratio $\exp(\beta)$.

TABLE 8.2. Data on Success of U.S. Government in Providing Health Care and Protecting the Environment

Health Care	Environment		
	Successful	Mixed	Unsuccessful
Successful	199	81	83
Mixed	129	167	112
Unsuccessful	164	169	363

Source: 2006 General Social Survey.

8.2.2 Example: Health Care and Environmental Protection

Table 8.2 summarizes responses to the questions "How successful is the government in (1) Providing health care for the sick? (2) Protecting the environment?" To detect whether responses tend to be more positive for one question than the other, we compare the marginal distributions using the cumulative logit model (8.2). The sample cumulative marginal proportions are (0.247, 0.526, 1.0) for health care and (0.335, 0.620, 1.0) for the environment. This shows a stochastic ordering, with responses on the environment tending more toward the low end of the ordinal scale (i.e., more successful) than those on health care. With scores (1, 2, 3), the mean response for health care was 2.23 and the mean response for environment was 2.04.

The cumulative logit model (8.2) has ML estimate $\hat{\beta} = 0.403$ (SE $= 0.058$).[1] There is strong evidence that population responses are more positive on the environment than on health care, with $z = (\hat{\beta}/\text{SE}) = 6.9$. The profile likelihood confidence interval for β is (0.290, 0.516). The fit of the model has $X^2 = 0.39$ (df $= 1$), a good fit. The estimated common cumulative odds ratio for comparing the two marginal distributions is $\exp(0.403) = 1.50$. For example, the estimated odds of response "successful" (instead of "mixed" or "unsuccessful") on the environment for a randomly selected subject are 1.50 times the estimated odds of the response "successful" on health care for another randomly selected subject.

8.2.3 Subject-Specific and Population-Averaged Tables

A three-way representation of the matched-pairs data motivates a different type of model. This display presents the data as n separate $2 \times c$ partial tables, one table for each matched pair. Partial table i shows the responses (Y_1, Y_2) for matched pair i, and we denote those responses by (Y_{i1}, Y_{i2}). Each table has columns that are the c possible outcomes for each observation. It shows the outcome for Y_{i1} in row 1 and the outcome for Y_{i2} in row 2.

Table 8.2 cross-classified results about government performance in health care and the environment for 1467 subjects. Table 8.3 shows a partial table for a subject who answers "successful" on health care and "mixed" on the environment. The full three-way table corresponding to Table 8.2 has 1467 partial tables. For example, 81 partial tables look like Table 8.3. Each subject has a partial table, displaying the two

---

[1]Results obtained using the R function mph.fit discussed in the Appendix.

TABLE 8.3. Representation of Matched Pair Contributing to Count n_{12} in Table 8.2

Issue	Response		
	Successful	Mixed	Unsuccessful
Health care	1	0	0
Environment	0	1	0

matched observations. The 1467 subjects provide 2934 observations in a $2 \times 3 \times 1467$ contingency table. Collapsing this table over the 1467 partial tables yields a 2×3 table with the first row equal to (363, 408, 696) and the second row equal to (492, 417, 558). These are the total number of (successful, mixed, unsuccessful) outcomes for the two questions and are the marginal counts in Table 8.2.

The $2 \times c \times n$ table having a separate partial table for the two responses by each of n subjects is called the *subject-specific table*. When each matched pair is not a single subject but, rather, a pair of matched subjects, such as in the occupational mobility study of Table 8.1, we refer to the table instead using the more general term *cluster-specific table*. A matched pair is a simple type of cluster. Models for the subject-specific table, such as the one in Section 8.2.4, are called *subject-specific models*, or *cluster-specific* in the more general case. Sometimes they are also referred to as *conditional models*, because when the data are stratified by subject, the effect comparing the responses is *conditional* on the subject. By contrast, the $c \times c$ table that cross-classifies in a *single* two-way table the two responses for *all* subjects is called a *population-averaged table*. Table 8.2 is an example. Its margins, which form the $2 \times c$ table obtained by collapsing the subject-specific tables, can yield estimates of population marginal probabilities. Models for the marginal probabilities, such as model (8.2), are called *marginal models*. In Chapters 9 and 10 we present marginal models and subject-specific models in detail, generalizing the models of this chapter by also permitting explanatory variables.

8.2.4 Subject-Specific Model Comparing Ordinal Responses

A subject-specific model for matched-pairs data refers to the subject-specific partial tables of form Table 8.3. A subject-specific cumulative logit model of proportional odds form is

$$\text{logit } [P(Y_{i1} \leq j)] = \alpha_i + \gamma_j, \qquad \text{logit } [P(Y_{i2} \leq j)] = \alpha_i + \gamma_j + \beta. \qquad (8.3)$$

Equivalently, logit $P(Y_{it} \leq j) = \alpha_i + \gamma_j + \beta x_{it}$ with $x_{i1} = 0$ and $x_{i2} = 1$. The $\{\gamma_j\}$ are monotone increasing in j. The model assumes that for each matched pair, the odds that observation 2 falls in category j or below (instead of above category j) are $\exp(\beta)$ times the odds for observation 1. Since each partial table refers to a single subject (matched pair), this conditional association is a *subject-specific effect*. When $\beta = 0$, the model states that for each matched pair, the response distribution is the same for both observations. This implies marginal homogeneity when averaged over all subjects.

This model differs from other models studied so far by permitting each subject to have their own probability distribution, reflected by the α_i term with the subject i subscript. The model permits subject heterogeneity, with each cumulative probability j having a value that varies among subjects. For identifiability, we impose a constraint, such as $\sum \alpha_i = 0$. A subject with a relatively large positive α_i has a relatively high probability of observation in category j or below for each of the two matched observations. By contrast, a subject with a relatively large negative α_i has a relatively low probability of observation in category j or below for each of the two matched observations.

Given the parameter values, model fitting treats the observations (Y_{i1}, Y_{i2}) for each matched pair as independent. Although the observations are treated as conditionally independent at the subject level, averaged over the subjects the model implies a nonnegative marginal association. The greater the variability in the $\{\alpha_i\}$, the greater the positive association between the two observations when considered marginally, because of the effect of the variation in value of α_i described in the preceding paragraph.

Statistical inference for the model usually focuses on parameter β for comparing the distributions. However, an awkward aspect is that the model has as many subject parameters $\{\alpha_i\}$ as subjects. This causes difficulties with the fitting process and with the properties of ordinary ML estimators. In Chapter 10 we discuss ML fitting of an extended model that treats $\{\alpha_i\}$ as *random effects*. This approach assumes that $\{\alpha_i\}$ are an unobserved sample having a particular probability distribution, such as the normal. Here, we present a simple estimate of β that although not ML for the random effects version of this model, is similar to that estimate and has similar SE. It is based on the fact that for each possible binary collapsing of the response to (category a and below, above category a), $a = 1, \ldots, c - 1$, the ratio of cell counts

$$\log \frac{\sum_{i > a} \sum_{j \leq a} n_{ij}}{\sum_{i \leq a} \sum_{j > a} n_{ij}}$$

estimates β. This is based on a standard result for the special case of binary matched pairs. Combining this information for all $c - 1$ possible collapsings yields an estimate proposed by Agresti and Lang (1993a),

$$\hat{\beta} = \log \frac{\sum \sum_{i > j} (i - j) n_{ij}}{\sum \sum_{j > i} (j - i) n_{ij}}. \tag{8.4}$$

McCullagh (1977) presented alternative estimates.

An ordinal test of marginal homogeneity ($\beta = 0$) can use this estimated effect (8.4) with its estimated standard error,

$$\mathrm{SE} = \sqrt{\frac{\sum \sum_{i < j} (j - i)^2 n_{ij}}{\left[\sum \sum_{i < j} (j - i) n_{ij}\right]^2} + \frac{\sum \sum_{i > j} (i - j)^2 n_{ij}}{\left[\sum \sum_{i > j} (i - j) n_{ij}\right]^2}}.$$

The ratio $z = \hat{\beta}/\mathrm{SE}$ is an approximate standard normal null test statistic.

8.2.5   Example: Health Care and the Environment Revisited

For Table 8.2 on government performance in providing health care and in protecting
the environment, Section 8.2.2 used a marginal cumulative logit model to compare
responses. That model had $\hat{\beta} = 0.403$ with SE $= 0.058$, for a common cumulative
odds ratio of 1.50 comparing the marginal distributions.

For Table 8.2, the estimator (8.4) for the subject-specific model equals

$$\hat{\beta} = \log \frac{1(129 + 169) + 2(164)}{1(81 + 112) + 2(83)} = \log \frac{626}{359} = 0.556.$$

The estimated subject-specific cumulative odds ratio is $\exp(\hat{\beta}) = 626/359 = 1.74$.
For each subject the estimated odds of response "successful" (instead of "mixed"
or "unsuccessful") on the environment are 1.74 times the estimated odds of that
response for health care. The estimate $\hat{\beta} = 0.556$ has SE $= 0.079$. For H_0: $\beta = 0$,
$z = 0.556/0.079 = 7.0$ provides extremely strong evidence against the null hypoth-
esis of marginal homogeneity, which the marginal model in Section 8.2.2 also
provided.

8.2.6   Marginal Effects Versus Subject-Specific Effects

With the subject-specific model, we just estimated the cumulative odds ratio com-
paring the two response distributions by 1.74. By contrast, the cumulative odds
ratio estimate of 1.50 found in Section 8.2.2 refers to a marginal model that
focused on the margins of this table. Those margins are equivalently the rows
of the marginal table obtained by collapsing the $2 \times 4 \times 1467$ contingency table
with subject-specific strata. That these odds ratios take different values reflects the
basic result that conditional odds ratios in three-way contingency tables can differ
substantially from marginal odds ratios in collapsed two-way tables.

As is the case with binary data, estimates of effects in subject-specific mod-
els tend to be larger in absolute value than estimates of corresponding effects in
marginal models. Compare, for example, $\hat{\beta} = 0.556$ just obtained with $\hat{\beta} = 0.403$
obtained for the marginal model. However, SE values also tend to be larger for
the subject-specific model, and usually $\hat{\beta}/SE$ takes a similar size. For Table 8.2,
$\hat{\beta}/SE = 6.9$ for the marginal model and 7.0 for the subject-specific model. In
Section 8.4.2 and Sections 10.1.4 and 10.5.1 in Chapter 10 we discuss further
the distinction between subject-specific and marginal models and their odds ratio
effects.

8.3   MODELS FOR THE JOINT DISTRIBUTION IN A SQUARE TABLE

An alternative analysis of square contingency tables models the joint distribution
of (Y_{i1}, Y_{i2}). Some such models have marginal homogeneity as a special case and
can also be used to compare the marginal distributions.

8.3.1 Symmetry Model for Matched Pairs

A $c \times c$ joint distribution $\{\pi_{ij}\}$ satisfies *symmetry* if

$$\pi_{ij} = \pi_{ji} \qquad \text{whenever } i \neq j. \tag{8.5}$$

Under symmetry, $\pi_{i+} = \sum_j \pi_{ij} = \sum_j \pi_{ji} = \pi_{+i}$ for all i, so marginal homogeneity also occurs. For $c = 2$, symmetry is equivalent to marginal homogeneity, but for $c > 2$ marginal homogeneity can occur without symmetry.

The ML fit for the symmetry model is $\hat{\pi}_{ij} = (n_{ij} + n_{ji})/2n = (p_{ij} + p_{ji})/2$ for all i and j. Its residual df $= c(c-1)/2$ for testing goodness of fit.

8.3.2 Ordinal Quasi-Symmetry Model for Matched Pairs

When the marginal distributions differ substantially, the symmetry model fits poorly. Marginal heterogeneity can be accommodated by the *quasi-symmetry model*. For matched-pairs data, this model has form

$$\log \frac{\pi_{ij}}{\pi_{ji}} = \beta_j - \beta_i \qquad \text{for all } i \text{ and } j, \tag{8.6}$$

with a constraint such as $\beta_c = 0$. Symmetry is the special case in which $\beta_1 = \cdots = \beta_{c-1} = 0$ also. The higher $\hat{\beta}_i$ compared to other $\{\hat{\beta}_j\}$, relatively more observations fall in column i than in row i.

The symmetry (S) and quasi-symmetry (QS) models treat the classifications as nominal. A special case of QS called *ordinal quasi-symmetry* (OQS) is often useful when the categories are ordered. Let $u_1 \leq u_2 \leq \cdots \leq u_c$ denote ordered scores for the rows and columns. The OQS model for matched-pairs data is

$$\log \frac{\pi_{ij}}{\pi_{ji}} = \beta(u_j - u_i). \tag{8.7}$$

This is a special case of the QS model in which $\{\beta_i\}$ have a linear trend. This model has the form of the usual logistic model, logit $\pi = \alpha + \beta x$, with $\alpha = 0$, $x = u_j - u_i$, and π equal to the conditional probability for cell (i, j), given response in cell (i, j) or cell (j, i). The greater the value of $|\beta|$, the greater the relative difference between π_{ij} and π_{ji} and between the marginal distributions. With scores $\{u_i = i\}$, the probability that the second observation is x categories higher than the first observation equals $\exp(x\beta)$ times the probability that the first observation is x categories higher than the second observation. An underlying bivariate normal latent model implies a model of this form (Agresti 1983, Exercise 8.1).

For sample cell proportions $\{p_{ij} = n_{ij}/n\}$ and sample marginal means $\sum_i u_i p_{i+}$ and $\sum_j u_j p_{+j}$, the likelihood equations for the ordinal quasi-symmetry model are

$$\sum_i u_i \hat{\pi}_{i+} = \sum_i u_i p_{i+}, \qquad \sum_j u_j \hat{\pi}_{+j} = \sum_j u_j p_{+j},$$

$$\hat{\pi}_{ij} + \hat{\pi}_{ji} = p_{ij} + p_{ji} \qquad \text{for } i < j.$$

The first two equations show that the fitted marginal distributions have the same means as the sample marginal distributions. When responses in one margin tend to be higher on the ordinal scale than those in the other margin, the fit of the OQS model (8.7) exhibits this same ordering. When $\hat{\beta} > 0$, the mean response is lower for the row variable. When $\hat{\beta} < 0$, the mean response is higher for the row variable. The ordinary QS model replaces the likelihood equations equating fitted and sample means with likelihood equations having the stronger condition that the fitted marginal proportions equal the sample marginal proportions. For each model we need only the equation for the rows or the equation for the columns, as the other is redundant because of the third equation above.

It is simple to fit the OQS model (8.7) using software for logistic models: Identify (n_{ij}, n_{ji}) as binomial numbers of successes and failures in $n_{ij} + n_{ji}$ trials, and fit a binary model with logit link function, with intercept forced to equal 0 (which most software can do with a "no intercept" option), and with the value of the predictor x equal to $u_j - u_i$ (Table A.5 in the Appendix illustrates). Since the OQS model has only one more parameter than the symmetry model, its residual df $= c(c-1)/2 - 1$ for testing goodness of fit. Symmetry and thus marginal homogeneity is the special case $\beta = 0$.

8.3.3 Example: Health Care and the Environment Revisited

We analyzed Table 8.2 on government performance in providing health care and in protecting the environment with cumulative logit models of marginal form in Section 8.2.2 and subject-specific form in Section 8.2.5. A cursory glance at the data reveals that the symmetry model is inappropriate. Indeed, $G^2 = 49.77$ for testing its fit, with df $= 3$. By comparison, the quasi-symmetry model fits well, having $G^2 = 0.72$ with df $= 1$. The simpler ordinal quasi-symmetry model also fits well. For the scores $\{1, 2, 3\}$, $G^2 = 0.76$ with df $= 2$, and $\hat{\beta} = 0.374$ (SE $= 0.055$). Table 8.4 displays its fitted values. The estimated probability that response on the environment is x categories more positive than the response on health care equals $\exp(0.374x)$ times the reverse probability. Responses on the environment tend to be more positive than those on health care.

Based on the fits of the S and OQS models, the likelihood-ratio test of marginal homogeneity uses the difference between the G^2 (deviance) values for the S and OQS models, with df $= 1$. This equals $49.77 - 0.76 = 49.0$. Alternatively, the Wald statistic $(\hat{\beta}/SE)^2 = (0.374/0.055)^2 = 46.6$. A third ordinal test is a score-type test, with chi-squared test statistic that is the square of the statistic discussed at the

TABLE 8.4. Fit of Ordinal Quasi-Symmetry Model to Table 8.2

y_1 = Health Care	y_2 = Environment		
	Successful	Mixed	Unsuccessful
Successful	199 (199.0)	81 (85.6)	83 (79.4)
Mixed	129 (124.4)	167 (167.0)	112 (114.6)
Unsuccessful	164 (167.6)	169 (166.4)	363 (363.0)

end of Section 8.1.2 that compares marginal sample means (for category scores $\{u_i = i\}$) relative to a null standard error. For these data, $z = (\bar{y}_1 - \bar{y}_2)/SE_0 = (2.227 - 2.045)/0.0262 = 6.9$, and $z^2 = 48.2$ with df $= 1$. All three statistics give very strong evidence of heterogeneity ($P < 0.0001$). The evidence is similar to that obtained in the preceding section for effects in a marginal model and in a subject-specific model, both of which had ($\hat{\beta}/SE$) values of about 7.

8.3.4 Diagonals-Parameter Symmetry Models

Goodman (1972, 1979b,c, 1981c, 1985) formulated several ordinal models in which diagonals that are parallel to and equidistant from the main diagonal exhibit similar patterns for probabilities. His *diagonals-parameter symmetry model* is

$$\log \frac{\pi_{ij}}{\pi_{ji}} = \beta_{j-i}, \qquad i < j, \tag{8.8}$$

for a set of parameters $\{\beta_1, \ldots, \beta_{c-1}\}$. The parameter β_k is the log odds that an observation falls in a cell (i, j) satisfying $j - i = k$ instead of in a cell satisfying $j - i = -k$, $k = 1, \ldots, c - 1$. Each probability on the diagonal that is k bands above the main diagonal is the same multiple $\exp(\beta_k)$ of the corresponding probability on the diagonal that is k bands below the main diagonal. The odds depend only on the distance k between the diagonal containing the cell and the main diagonal.

The symmetry model is the special case of this model in which $\beta_1 = \cdots = \beta_{c-1} = 0$. The diagonals-parameter symmetry model has $c - 1$ more parameters than the symmetry model, and its residual df $= (c - 1)(c - 2)/2$. The ordinal quasi-symmetry model is the special case of (8.8) that replaces $\{\beta_k\}$ by a single parameter, through

$$\beta_{j-i} = (u_j - u_i)\beta.$$

Let $\mathbf{n}_k$ denote the $2 \times (c - k)$ table constructed using the two diagonals that are k bands from the main diagonal, $k = 1, \ldots, c - 2$. For example, $\mathbf{n}_1$ has first row $(n_{12}, n_{23}, \ldots, n_{(c-1)c})$ and second row $(n_{21}, n_{32}, \ldots, n_{c(c-1)})$. When the diagonals-parameter symmetry model holds, the expected values of these counts are such that the entry in the first row of $\mathbf{n}_k$ is the same multiple $\exp(\beta_k)$ of the corresponding entry in the second row. The model is therefore equivalent to independence for the expected frequencies in the tables $\{\mathbf{n}_k, k = 1, \ldots, c - 2\}$. The fitted values $\{\hat{\mu}_{ij}\}$ for the model are the expected frequency estimates for the independence model applied to the $\{\mathbf{n}_k\}$ tables. The estimate of β_k is the common log of the within-column ratio of expected frequency estimates in the two rows of $\mathbf{n}_k$.

8.3.5 Conditional Symmetry Model

Bishop et al. (1975, pp. 285–286) proposed the model

$$\log \frac{\pi_{ij}}{\pi_{ji}} = \beta, \qquad i < j. \tag{8.9}$$

This model implies that for $i < j$,

$$P(Y_1 = i, Y_2 = j \mid Y_1 < Y_2) = P(Y_1 = j, Y_2 = i \mid Y_1 > Y_2),$$

where (Y_1, Y_2) is selected at random according to the $\{\pi_{ij}\}$ distribution. Because of this property, it is referred to as a *conditional symmetry model*. This model is the special case of the diagonals-parameter symmetry model (8.8) with $\beta_1 = \cdots = \beta_{c-1} = \beta$. The symmetry model corresponds to $\beta = 0$. Like the ordinal quasi-symmetry model, the conditional symmetry model has only one more parameter than the symmetry model, and its residual df $= c(c-1)/2 - 1$.

The likelihood equations for the conditional symmetry model are

$$\hat{\mu}_{ij} + \hat{\mu}_{ji} = n_{ij} + n_{ji} \quad \text{for all } i \leq j, \qquad \sum\sum_{i<j} \hat{\mu}_{ij} = \sum\sum_{i<j} n_{ij}.$$

The ML estimate of β has closed form,

$$\hat{\beta} = \log \frac{\sum\sum_{i<j} n_{ij}}{\sum\sum_{i>j} n_{ij}}.$$

This estimate has estimated asymptotic variance

$$\left(\sum\sum_{i<j} n_{ij}\right)^{-1} + \left(\sum\sum_{i>j} n_{ij}\right)^{-1}.$$

The fitted values for the model are

$$\hat{\mu}_{ij} = \frac{\exp(\hat{\beta})(n_{ij} + n_{ji})}{\exp(\hat{\beta}) + 1}, \quad i < j \quad \text{and}$$

$$\hat{\mu}_{ij} = \frac{n_{ij} + n_{ji}}{\exp(\hat{\beta}) + 1}, \quad i > j,$$

with $\hat{\mu}_{ii} = n_{ii}$ for $i = 1, \ldots, c$.

8.3.6 Quasi-Independence and Quasi-Uniform Association

Often, regular loglinear models for ordinal variables fit square tables well when they are adapted to fit the cells on the main diagonal perfectly, as the conditional symmetry model does. Let $\{u_i\}$ be ordered category scores. Consider the loglinear model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i u_j + \delta_i I(i = j), \tag{8.10}$$

where $I(i = j)$ denotes an indicator that equals 1 when $i = j$ and 0 otherwise. This model adds the main diagonal parameters $\{\delta_i\}$ to the linear-by-linear association model (6.2): namely, δ_1 for cell (1,1) (in row 1 and column 1), δ_2 for cell (2,2), and so on. The ML fit in those cells is perfect, with $\hat{\mu}_{ii} = n_{ii}$ for all i. The model permits linear-by-linear association off the main diagonal. This model is called the *quasi-linear-by-linear association model*.

For equal-interval scores, model (8.10) implies uniform local association, given that responses differ. In that case, Goodman (1979a) called the model *quasi-uniform association*. See also Duncan (1979). Model (8.10) has residual df $= c(c - 3)$, so it is unsaturated only for tables of size 4×4 and larger. The fit can be obtained with standard software for GLMs, as Table A.5 in the Appendix shows. A more general model treats the scores as parameters, as did the RC model of Section 6.5.1. A simpler model introduced in Section 8.5.3 sets $\delta_1 = \cdots = \delta_r$.

The special case $\beta = 0$ of model (8.10) is the *quasi-independence model*, an important model for square tables with unordered categories. When $\delta_i > 0$ in that model, cell (i, i) is more likely than under independence. Off the main diagonal, the ordinary independence model applies. In other words, conditional on the observations differing, Y_1 is independent of Y_2. For 3×3 tables, it is equivalent to the quasi-symmetry model.

8.3.7   Example: Health Care and the Environment Revisited

We analyzed Table 8.2 on opinions about government spending on health care and on the environment in Sections 8.2.2, 8.2.5, and 8.3.3 using various ordinal models. Section 8.3.3 found that the ordinal quasi-symmetry model fits well, so necessarily the diagonals-parameter symmetry model (8.8), which is more complex, fits well. Its deviance statistic is $G^2 = 0.08$ (df $= 1$) and its parameter estimates are $\hat{\beta}_1 = -0.434$ and $\hat{\beta}_2 = -0.681$. For example, the estimated odds that response on the environment is one category more positive than response on health care (instead of the reverse) is $\exp(-\hat{\beta}_1) = \exp(0.434) = 1.54$. For this model, $\exp(-\hat{\beta}_2) = 1.98 = n_{31}/n_{13} = 164/83$.

The conditional symmetry model (8.9) also fits well, with deviance $G^2 = 2.39$ (df $= 2$). It has estimate $\hat{\beta} = -0.515$ (SE $= 0.076$). The estimated odds that response on the environment is more positive than response on health care (instead of the reverse) equals $\exp(0.515) = 1.67$. Finally, the quasi-independence model also fits well ($G^2 = 0.72$, df $= 1$), being equivalent to the quasi-symmetry model for these data.

8.4   COMPARING MARGINAL DISTRIBUTIONS FOR MATCHED SETS

The methods for comparing marginal distributions of matched pairs extend to marginal distributions of matched sets. To show this, we first extend the notion of marginal homogeneity from square tables to T-dimensional tables and express it is a special case of marginal models and cluster-specific models.

8.4.1 Marginal Homogeneity in Tables for Matched Sets

Let $Y_1, Y_2, \ldots, Y_T$ denote the T responses in a matched set. With c response categories, a contingency table with c^T cells summarizes the possible sequences of observations in a cluster. Let $\mathbf{i} = i_1, \ldots, i_T$ denote the cell having $Y_t = i_t$, $t = 1, \ldots, T$. Let $\pi_{\mathbf{i}} = P(Y_t = i_t, t = 1, \ldots, T)$. Then

$$P(Y_t = j) = \pi_{+ \cdots + j + \cdots +},$$

where the j subscript is in position t, and $\{P(Y_t = j), j = 1, \ldots, c\}$ is the marginal distribution for Y_t.

This T-way table satisfies *marginal homogeneity* if

$$P(Y_1 = j) = P(Y_2 = j) = \cdots = P(Y_T = j), \qquad \text{for } j = 1, \ldots, c - 1.$$

As in the matched-pairs case, various ordinal models have marginal homogeneity as a special case. Having $df = T - 1$ for tests comparing margins, these models provide more powerful ways of detecting shifts in location that standard (nominal-scale) analyses that have $df = (T - 1)(c - 1)$.

8.4.2 Marginal and Subject-Specific Models for Matched Sets

As in Section 8.2, we consider both marginal models and subject-specific models. A marginal cumulative logit model with proportional odds structure for the T margins of the c^T contingency table is

$$\text{logit } [P(Y_t \le j)] = \alpha_j + \beta_t, \qquad j = 1, \ldots, c - 1, \quad t = 1, \ldots, T, \qquad (8.11)$$

with a constraint such as $\beta_T = 0$. The odds of response below any particular category for a subject randomly selected for observation t equal $\exp(\beta_t - \beta_u)$ times the odds for another subject randomly selected for observation u. Marginal homogeneity is the special case $\beta_1 = \cdots = \beta_T$.

Although model (8.11) is simple, ML fitting of it is not straightforward. As Section 8.2.1 explained in the matched-pairs setting, the model refers to *marginal probabilities*, whereas the multinomial likelihood function uses the *joint distribution* of the data. ML fitting can be accomplished with special-purpose programs such as the mph.fit R function described in the Appendix.

Conditional, subject-specific models for matched pairs also extend to matched sets. Let Y_{it} denote observation t in cluster i. A cumulative logit model with proportional odds structure is

$$\text{logit } [P(Y_{it} \le j)] = \alpha_i + \gamma_j + \beta_t, \qquad j = 1, \ldots, c - 1, \quad t = 1, \ldots, T. \qquad (8.12)$$

For cluster i, the odds of response below any particular category for observation t equal $\exp(\beta_t - \beta_s)$ times the odds for observation s. Marginal homogeneity is

implied by $\beta_1 = \cdots = \beta_T$. The same interpretation applies to a more general model that replaces $\alpha_i + \gamma_j$ by α_{ij}, letting the distances between cutpoints vary by cluster.

For $c = 2$, a model of form (8.12) is called the *Rasch model*. This model is often used in educational applications as an *item response model*, for example to model the probability that subject i makes a correct response on question t, using subject parameters $\{\alpha_i\}$ and item (question) parameters $\{\beta_t\}$. For $c > 2$, Samejima (1969) considered such models for cumulative logits and probits.

As in the matched-pairs case, the complication with this model is the large number of cluster-specific parameters $\{\alpha_i\}$. To reduce the number of parameters, the random effects approach treats $\{\alpha_i\}$ as unobserved random variates having distribution in a parametric family, such as the normal. In Chapter 10 we study models such as (8.12) in more general contexts, incorporating both matched sets and explanatory variables.

8.4.3 Example: Crossover Study for Treating Dysmenorrhea

Table 8.5, based on data from Kenward and Jones (1991), shows results of a three-period crossover study. The study was designed to compare placebo (treatment A) with a low-dose analgesic (treatment B) and a high-dose analgesic (treatment C) for relief of severe uterine pain during a woman's menstrual cycle, a condition called *dysmenorrhea*. The study assigned subjects randomly to one of the six possible sequences for administering the three treatments during the three periods. At the end of each period, each woman rated the treatment as giving no relief (1), moderate relief (2), or complete relief (3). More complex modeling showed no evidence that effects depended on the sequence, and Table 8.5 shows the data collapsed over the six sequences.

We consider the marginal cumulative logit model

$$\text{logit } P(Y_t \le j) = \alpha_j + \beta_t, \qquad j = 1, 2, \quad t = A, B, C, \qquad (8.13)$$

with constraint $\beta_C = 0$. We report here estimates of $\{\beta_t\}$ for the reparameterization with $P(Y_t \le j)$ replaced by $P(Y_t > j)$, so that larger values of β_t correspond to

TABLE 8.5. Data from Crossover Study for Treating Dysmenorrhea, with Fitted Values for Ordinal Quasi-Symmetry Model[a]

		B = 1			B = 2			B = 3		
A	C:	1	2	3	1	2	3	1	2	3
1		6	4	5	3	13	10	1	8	14
		(6.0)	(4.7)	(4.5)	(3.3)	(11.2)	(9.6)	(2.3)	(6.9)	(15.0)
2		2	3	2	1	3	1	2	1	2
		(1.0)	(3.4)	(2.9)	(2.4)	(3.0)	(1.0)	(1.5)	(0.7)	(2.0)
3		1	0	2	0	0	0	1	1	0
		(0.2)	(0.6)	(1.3)	(0.4)	(0.2)	(0.6)	(0.7)	(0.4)	(0.0)

Source: Kenward and Jones (1991).
[a] 1, giving no relief; 2, moderate relief; 3, complete relief.

more positive results (e.g., a greater probability of complete relief). Equivalently, such estimates result from replacing β_t in the model by $-\beta_t$, which in Section 3.3.2 we noted is the natural parameterization induced by an underlying latent variable model. The ML estimates[2] are then

$$\hat{\beta}_B - \hat{\beta}_A = 2.038 \qquad (SE = 0.360),$$

$$\hat{\beta}_C - \hat{\beta}_A = 2.430 \qquad (SE = 0.372),$$

$$\hat{\beta}_C - \hat{\beta}_B = 0.392 \qquad (SE = 0.252).$$

For example, the estimated odds that relief for the low-dose analgesic is complete or moderate rather than none, or complete rather than moderate or none, are $\exp(2.038) = 7.7$ times the estimated odds for the placebo. Treatments B and C clearly differ from placebo, but there is only weak evidence that the high dose works better than the low dose.

ML model fitting provides fitted values in the cells of the $3{\times}3{\times}3$ table that satisfy the model. Those fitted values can be compared to the observed counts using chi-squared statistics to test the fit of the model. For these data, $X^2 = 0.5$ and $G^2 = 0.5$ (df $= 2$) for comparing observed to fitted cell counts in modeling the six marginal logits (two for each treatment) using the four parameters of model (8.13). The fit is quite good. The likelihood-ratio test of H_0: $\beta_A = \beta_B = \beta_C$ has a test statistic that equals the difference between the deviance for the simpler model imposing this constraint and the deviance for model (8.13). It equals $46.3 - 0.5 = 45.8$ with df $= 2$, strong evidence of at least one difference among the three treatments.

8.4.4 Ordinal Quasi-Symmetry Model for Matched Sets

In Section 8.3 we showed that some models for the joint distribution of matched pairs have parameters that compare the marginal distributions. Some of these models, such as the ordinal quasi-symmetry model that has marginal means in its set of sufficient statistics, extend to matched sets. For a multinomial sample of size n, let $\mu_i = n\pi_i$ denote the expected frequency in cell i.

The joint distribution for a c^T contingency table satisfies *complete symmetry* if $\pi_i = \pi_j$ for any permutation $j = j_1, \ldots, j_T$ of $i = i_1, \ldots, i_T$. Complete symmetry can be expressed as the loglinear model

$$\log \mu_i = \lambda_{ab\ldots m},$$

where a is the minimum of $i_1, \ldots, i_T$, b is the next smallest, $\ldots$, and m is the maximum. This notation reflects the permutation invariance of μ_i in the subscript i. In a three-way table, for example, $\log \mu_{133} = \log \mu_{313} = \log \mu_{331} = \lambda_{133}$.

---

[2]Found using the mph.fit R function discussed in the Appendix as shown at www.stat.ufl.edu/~aa/ordinal/ord.html.

A c^T contingency table cross-classifying $Y_1, Y_2, \ldots, Y_T$ satisfies *quasi-symmetry* (QS) if

$$\log \mu_{\mathbf{i}} = \lambda + \lambda_{i_1}^{Y_1} + \lambda_{i_2}^{Y_2} + \cdots + \lambda_{i_T}^{Y_T} + \lambda_{ab\ldots m}, \tag{8.14}$$

where $\lambda_{ab\ldots m}$ has the permutation invariant structure of the complete symmetry model. This model permits each single-factor marginal distribution to have its own parameters. The model's likelihood equations imply that the fitted marginal distributions are identical to the sample marginal distributions.

For ordinal responses, a simpler loglinear model with quantitative main effects uses ordered scores $\{u_j\}$ for the c categories in each margin. The *ordinal quasi-symmetry* (OQS) model is

$$\log \mu_{\mathbf{i}} = \beta_1 u_{i_1} + \beta_2 u_{i_2} + \cdots + \beta_T u_{i_T} + \lambda_{ab\ldots m}, \tag{8.15}$$

where $\lambda_{ab\ldots m}$ is permutation invariant and $\{\beta_t\}$ satisfy a constraint such as $\beta_T = 0$. For matched pairs, in logit form for cells (i, j) and (j, i), the equivalent expression is

$$\log \frac{\mu_{ij}}{\mu_{ji}} = \beta(u_j - u_i) \qquad \text{for all } i \text{ and } j,$$

identifying β with $\beta_2 - \beta_1$ in model (8.15) for $T = 2$. This equation corresponds to equation (8.7) in terms of probabilities.

For the OQS model, complete symmetry is the special case $\beta_1 = \cdots = \beta_T$. The sufficient statistic for β_k is the mean response for sample marginal distribution k. The sufficient statistic for the $\lambda_{ab\ldots m}$ term is the sum of all cell counts having those indices. For a three-way table, for example, the sufficient statistic for λ_{133} is $n_{133} + n_{313} + n_{331}$. The likelihood equations equate the sufficient statistics to their fitted values. The complete symmetry model has residual df $= c^T - \binom{c+T-1}{T}$, the OQS model has residual df $= c^T - \binom{c+T-1}{T} - (T - 1)$, and the ordinary QS model (8.14) has residual df $= c^T - \binom{c+T-1}{T} - (c - 1)(T - 1)$.

Since their sufficient statistics are the marginal means, the $\{\beta_k\}$ in the OQS model (8.15) reflect shifts in location among the T marginal distributions. Like the marginal cumulative logit model (8.11), this model is useful for describing location shifts in the marginal distributions. Both models fit poorly when the margins have quite different dispersion. The ML estimates $\{\hat{\beta}_j\}$ for OQS have the same order as the sample mean responses in the marginal distributions.

When the OQS model holds, marginal homogeneity is equivalent to complete symmetry (S). Marginal heterogeneity occurs if OQS holds but S does not. The likelihood-ratio statistic comparing the deviances tests marginal homogeneity, with df $= T - 1$. This test is sensitive to detecting location shifts in the marginal means. The score test is based on variability among those means, as discussed in Section 8.1.2 for $T = 2$ and in Section 8.4.8 for general T.

8.4.5 Ordinal Quasi-Symmetry and Ordinal Rasch Models

A useful way to interpret $\{\beta_k\}$ in the ordinal quasi-symmetry model (8.15) utilizes a connection with a cluster-specific adjacent-categories logit model,

$$\log \frac{P(Y_{it} = j + 1)}{P(Y_{it} = j)} = \alpha_{ij} + \beta_t, \qquad j = 1, \ldots, c - 1, \quad t = 1, \ldots, T. \quad (8.16)$$

One way to estimate $\{\beta_t\}$ treats $\{\alpha_{ij}\}$ as fixed effects and conditions on their sufficient statistics to eliminate them from the likelihood function. Then the conditional ML estimates of $\{\beta_t\}$ for model (8.16) are identical to the ordinary ML estimates of $\{\beta_t\}$ for the OQS model (8.15) with $\{u_j = j\}$ (Agresti 1993a,b, 1995). So $\{\beta_t\}$ have the interpretation that for each given cluster i, the odds of the higher instead of the lower of two adjacent responses for observation t are $\exp(\beta_t - \beta_s)$ times the corresponding odds for observation s.

Model (8.16) is an extension of the *Rasch model* from a binary response for cluster i and "item" t to an ordinal response. A simpler extension also gives structure to the cluster-specific terms (Andersen 1973; Andrich 1978),

$$\log \frac{P(Y_{it} = j + 1)}{P(Y_{it} = j)} = \alpha_i + \gamma_j + \beta_t. \quad (8.17)$$

This linear predictor structure is the same as in the corresponding subject-specific cumulative logit model (8.12). The conditional ML estimates of $\{\beta_t\}$ for this ordinal Rasch model are identical to the ordinary ML estimates of $\{\beta_t\}$ for the special case of the OQS model (8.15),

$$\log \mu_i = \beta_1 u_{i_1} + \beta_2 u_{i_2} + \cdots + \beta_T u_{i_T} + \sum_{j=1}^{c} \lambda_j t_j + \rho_s \quad (8.18)$$

with $s = i_1 + i_2 + \cdots + i_T$, $\{u_j = j\}$, and t_j denotes the number of $\{i_t\}$ that equal j. Here the symmetry term ρ_s refers to a coarser partition having a separate parameter for each sum of responses rather than for each possible ordered sequence. This ordinal Rasch model has a simple form for the cluster and item effects for a given outcome category. When interest focuses mainly on $\{\beta_t\}$, the more general model (8.16) provides more flexibility, because the corresponding loglinear model (8.15) fits well in a wider variety of cases. For related models in an item response context, see Masters (1982).

In Section 8.3 we introduced several other models for square tables with ordered categories, such as diagonals-parameter-symmetry and conditional symmetry. These do not extend as simply to matched sets as the ordinal quasi-symmetry model does.

8.4.6 Example: Dysmenorrhea Crossover Study Revisited

We now further analyze Table 8.5, from a crossover study comparing three treatments for relief of dysmenorrhea. Table 8.6 shows the goodness of fit of several

TABLE 8.6. Goodness of Fit of Loglinear Models for Table 8.5

Model	Deviance G^2	Pearson X^2	Degrees of Freedom
Mutual independence	26.73	25.06	20
Complete symmetry	69.00	66.21	17
Ordinal quasi-symmetry	10.35	10.01	15
Quasi-symmetry	9.99	9.37	13

loglinear models. Because the table is sparse, we put more faith in the Pearson X^2 statistic, using the deviance mainly to compare models. The OQS model (8.15) fits well ($X^2 = 10.0$, df $= 15$). Table 8.5 also displays its fitted values. There is a dramatic improvement compared to the complete symmetry model (which has $X^2 = 66.2$, df $= 17$). The ordinary QS model fits only slightly better, and interpretations are simpler for the ordinal model.

Denote main effects in the ordinal quasi-symmetry model for the treatments by β_A, β_B, β_C. The likelihood-ratio test of H_0: $\beta_A = \beta_B = \beta_C$ has test statistic $G^2(S) - G^2(OQS) = 69.0 - 10.3 = 58.7$, with df $= 2$. The estimated treatment comparisons are:

$$\hat{\beta}_B - \hat{\beta}_A = 1.207 \quad (SE = 0.239),$$

$$\hat{\beta}_C - \hat{\beta}_A = 1.537 \quad (SE = 0.259),$$

$$\hat{\beta}_C - \hat{\beta}_B = 0.330 \quad (SE = 0.221).$$

Substantive results are the same as with the marginal cumulative logit model (Section 8.4.3). Treatments B and C both clearly differ from the placebo (A), but there is only weak evidence that the high dose is better than the low dose. These effects can also be interpreted in terms of the generalization (8.16) of the Rasch model. For instance, for a given subject, the estimated odds that relief for the low-dose analgesic is moderate rather than none, or complete rather than moderate, are $\exp(1.207) = 3.34$ times the estimated odds for the placebo.

8.4.7 Comparing Marginal Means for Matched Sets

For simpler interpretation, it can be helpful to report sample marginal means and their differences and SE values. With response scores (1, 2, 3) for the possible outcomes for each treatment for dysmenorrhea, the sample means were 1.31 for treatment A, 2.06 for B, and 2.22 for C. The differences between pairs of sample means, with SE values based on equation (8.1) for matched-pairs data, are

$$\bar{y}_B - \bar{y}_A = 0.744 \quad (SE = 0.113),$$

$$\bar{y}_C - \bar{y}_A = 0.907 \quad (SE = 0.109),$$

$$\bar{y}_C - \bar{y}_B = 0.163 \quad (SE = 0.105).$$

These results correspond to estimates from fitting the model

$$\sum_j v_j P(Y_t = j) = \beta_t, \qquad t = 1, 2, \ldots, T, \qquad (8.19)$$

where $\{v_j = j\}$. Unless we impose structure on $\{\beta_t\}$, such as a linear trend, the model is saturated. Here, the substantive results are the same as with models that account more fully for the categorical nature of the data.

The margins can be compared in significance tests using the mean responses. Koch et al. (1977) proposed Wald-type tests. The generalized CMH tests described next are score-type tests.

8.4.8 Generalized CMH Tests Comparing Marginal Means

In Section 6.4.5 we introduced generalized Cochran–Mantel–Haenszel (CMH) tests for testing conditional independence in three-way contingency tables. In this context of comparing the T margins of a c^T contingency table, we apply the test to cluster-specific partial tables that show the T observations for a particular cluster. That is, for cluster i, partial table i has rows that are the T observations and columns that are the c possible response outcomes. One observation occurs in each row. Such a data file corresponds naturally to the cluster-specific sort of model of form (8.12), which in Chapter 10 we generalize and study further.

One such generalized CMH test assigns scores to the response categories and is sensitive to detecting variability in the T marginal means. In this sense, it connects with the mean response model (8.19) just described. For details about such tests in this context, see White et al. (1982) and Landis et al. (1978, 1988). The covariance matrix of the sample marginal means is evaluated under the null hypothesis of equal means, so this is a score-type statistic for the model (8.19) for marginal means and also for the ordinal quasi-symmetry model, which has the marginal means as its sufficient statistics.

When we apply this approach with the crossover study using response scores (1, 2, 3), the generalized CMH statistic for comparing the means equals 73.4 with df = 2. Like the model-based analyses, this shows extremely strong evidence of a difference among the treatments.

8.5   ANALYZING RATER AGREEMENT ON AN ORDINAL SCALE

One application in which matched-pairs data occur is the investigation of agreement between two observers or "raters." For example, for each of several patients, two doctors might evaluate whether the patient has a particular condition, with a scale such as (yes, probably yes, probably no, no). Sometimes, one rater is considered an "expert." Then, interest focuses on how well another rater or set of raters tends to agree with the expert.

Table 8.7 illustrates a rater agreement application in which the raters were movie critics. The table summarizes ratings of 160 movies between April 1995

TABLE 8.7. Ratings of Movies by Gene Siskel and Roger Ebert[a]

Siskel Rating	Ebert Rating		
	Con	Mixed	Pro
Con	24 (24.0[*], 24.1[†])	8 (8.1, 6.3)	13 (12.9, 14.5)
Mixed	8 (7.9, 6.2)	13 (13.0, 14.6)	11 (11.1, 11.2)
Pro	10 (10.1, 11.6)	9 (8.9, 9.1)	64 (64.0, 62.3)

Source: Adapted from A. Agresti and L. Winner, *CHANCE* **10**: 10–14, (1997), with permission. Copyright © 1997 American Statistical Association. All rights reserved.
[a]The first entry (*) in each pair is the fit of the quasi-independence model; the second entry (†) is the fit of the simpler model, which adds a single main diagonal parameter to the independence model.

and September 1996 for at-the-time Chicago newspaper film critics Gene Siskel and Roger Ebert. Each movie review is categorized with the scale (pro, con, mixed) according to whether the review was positive, negative, or a mixture of the two. In Table 8.7, each matched pair consists of the ratings by Siskel and Ebert for a particular movie.

8.5.1 Agreement: Departures from Independence

For Table 8.7, let π_{ij} denote the probability that Siskel classified a movie in category i and Ebert classified it in category j. Their ratings of a particular movie *agree* if their classifications are in the same category. In a square table, the main diagonal $\{i = j\}$ represents rater agreement. Thus, $\sum_i \pi_{ii}$ is the total probability of agreement. Perfect agreement occurs when $\sum_i \pi_{ii} = 1$.

For cross-classifications of ordinal rater evaluations, we usually expect a positive association. In fact, the independence model fits poorly ($G^2 = 43.2$, df = 9). Its standardized residuals (not shown here) take large positive values on the main diagonal, indicating that agreement for each category is greater than expected if the ratings were statistically independent.

8.5.2 Quasi-Independence and Agreement Modeling

More complex models add components that relate to agreement beyond that expected under independence. For two raters A and B, a useful generalization is quasi-independence. For expected frequencies $\{\mu_{ij}\}$, this is the loglinear model

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \delta_i I (i = j),$$

which adds main-diagonal parameters $\{\delta_i\}$ to the independence model. For the movie reviewer data, this model has deviance $G^2 = 0.01$ (df = 1). It fits much better than the independence model. Table 8.7 shows the fit. The fitted counts have the same main-diagonal values and the same row and column totals as the observed data, but satisfy independence for cells not on the main diagonal. Given that the critics disagreed, the rating by Siskel seems to have been essentially independent of the rating by Ebert.

Conditional on rater disagreement on an ordinal scale, an association often remains. Thus, the quasi-symmetry model often fits much better than the quasi-independence model. With $c = 3$, however, these models are equivalent.

Models for agreement can take ordering of categories into account. The ordinal quasi-symmetry model (8.5) fits these data well, with $G^2 = 0.09$ (df = 2). The small value of $\hat{\beta} = 0.1255$ for that model (SE = 0.178) suggests that the simpler model of symmetry is adequate. In fact, this is a rare data set for which this is true ($G^2 = 0.59$, df = 3). However, these models do not have parameters that naturally summarize agreement. For example, the parameter β in the OQS model focuses on a location difference between the margins.

8.5.3 Ordinal Agreement Model

Conditional on rater disagreement, there is often a tendency for high (low) ratings by one rater to occur with relatively high (low) ratings by the other rater. More useful quasi-symmetric models for ordinal agreement describe this by partitioning beyond-chance agreement into that due to a baseline association and a main-diagonal increment. One such model that generalizes the quasi-independence model is the quasi linear-by-linear association model (8.10). Conditional on disagreement, there is a linear-by-linear association. This model has residual df = $c(c - 3)$, however, so it is saturated for Table 8.7.

A more parsimonious model (Agresti 1988) is the *ordinal agreement model*

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta I(i = j). \tag{8.20}$$

This uses a single parameter δ to describe beyond-chance agreement on the main diagonal and another parameter β to describe association off that diagonal. The likelihood equations equate the model fitted values to the data for the marginal distributions (the sufficient statistics for $\{\lambda_i^A\}$ and $\{\lambda_j^B\}$), the correlation for the chosen scores (the sufficient statistic for β), and the prevalence of exact agreement (the sufficient statistic for δ),

$$\sum_i \hat{\mu}_{ii} = \sum_i n_{ii}.$$

So this model accommodates the larger number of observations that typically fall on the main diagonal relative to what the independence model predicts. The model can be fitted with standard software for loglinear models (see Table A.5 in the Appendix).

For the movie reviewers' data, the ordinal agreement model with $\{u_i = i\}$ has deviance $G^2 = 0.59$ (df = 2). The agreement and association parameter estimates are $\hat{\delta} = 0.859$ (SE = 0.300) and $\hat{\beta} = 0.194$ (SE = 0.206). This suggests that a simpler model that adds only a single main diagonal parameter to the independence model fits well. In fact, that is the case, with fit statistics $G^2 = 1.49$ (df = 3) and with $\hat{\delta} = 1.094$ (SE = 0.171). Table 8.7 also shows this fit. The estimated odds of agreement for any given category are $\exp(1.094) = 3.0$ times what the independence model predicts.

8.5.4 Odds Ratio Summarizing Agreement

For two subjects observed by two raters, suppose that each rater classifies one in category i and one in category j. The odds that the raters agree rather than disagree on which is in category i and which is in category j equal

$$\tau_{ij} = \frac{\pi_{ii}\pi_{jj}}{\pi_{ij}\pi_{ji}} = \frac{\mu_{ii}\mu_{jj}}{\mu_{ij}\mu_{ji}}.$$

As τ_{ij} increases, the raters are more likely to agree for that pair of categories. For the ordinal agreement model (8.20), $\log \tau_{ij} = (u_j - u_i)^2\beta + 2\delta$. When $\beta > 0$ and $\delta > 0$, the agreement log odds are greater for categories farther apart. For instance, for Table 8.7, $\hat{\delta} = 0.859$ and $\hat{\beta} = 0.194$, so that $\hat{\tau}_{12} = \hat{\tau}_{23} = 6.8$, whereas $\hat{\tau}_{13} = 12.1$.

As often happens in practice, more than one model fits well for the movie reviewer data. It's not necessary to restrict attention to one model, as different models describe different aspects. For example, although the symmetry model and the quasi-independence model fit adequately, the ordinal agreement model is more useful for showing how the observations cluster around the main diagonal.

8.5.5 Weighted Kappa: A Summary Measure of Agreement

An alternative approach summarizes the agreement with a single index. For nominal scales, the most popular measure of agreement is *kappa* (Cohen 1960). It compares the agreement to that expected if the ratings were independent. Kappa treats classifications as nominal. So it treats a disagreement for ordered categories that are close the same as a disagreement for categories that are far apart.

When categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. The measure *weighted kappa* (Spitzer et al. 1967) uses weights $\{w_{ij}\}$ satisfying $0 \leq w_{ij} \leq 1$ with all $w_{ii} = 1$ and all $w_{ij} = w_{ji}$ to describe closeness of agreement. Popular choices for weights are

$$\left\{ w_{ij} = \frac{1 - |i - j|}{c - 1} \right\} \quad \text{and} \quad \left\{ w_{ij} = \frac{1 - (i - j)^2}{(c - 1)^2} \right\}.$$

For both of these, agreement is weaker and disagreement is stronger for cells farther from the main diagonal. The weighted agreement is defined to be $\sum_i \sum_j w_{ij}\pi_{ij}$. Weighted kappa compares this to its expected value under independence, using

$$\kappa_w = \frac{\sum_i \sum_j w_{ij}\pi_{ij} - \sum_i \sum_j w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum_i \sum_j w_{ij}\pi_{i+}\pi_{+j}}.$$

The denominator equals the numerator with the weighted agreement replaced by its maximum possible value of 1, corresponding to perfect agreement, $\sum_i \pi_{ii} = 1$. The ordinary kappa (unweighted) has $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$.

Weighted kappa equals 0 when the independence model holds, and it equals 1.0 when perfect agreement holds. The stronger the weighted agreement, the higher the

value of κ_w, for given marginal distributions. With $\{w_{ij} = 1 - (i - j)^2/(c - 1)^2\}$, Fleiss and Cohen (1973) showed that κ_w is an intraclass correlation for a two-way ANOVA treating the subjects rated and the raters as random samples of subjects and raters. The intraclass correlation is the ratio of the variability between subjects to the total variability between subjects and between raters. That is, like r-squared, it is a "proportion of variability explained" sort of measure.

For sample cell proportions $\{p_{ij}\}$, let

$$\bar{w}_{i.} = \sum_j w_{ij} p_{+j}, \qquad \bar{w}_{.j} = \sum_i w_{ij} p_{i+}$$

denote weighted averages of the weights in row i and column j, and let $P_o = \sum_i p_{ii}$ and $P_e = \sum_i p_{i+} p_{+i}$. The estimated asymptotic variance of the sample value $\hat{\kappa}_w$ is (Fleiss et al. 1969)

$$(SE)^2 = \frac{1}{n(1 - P_e)^4} \left\{ \sum_i \sum_j p_{ij} [w_{ij}(1 - P_e) - (\bar{w}_{i.} + \bar{w}_{.j})(1 - P_o)]^2 \right.$$

$$\left. - (P_o P_e - 2P_e + P_o)^2 \right\}.$$

For Table 8.7 for the movie reviewers, $\hat{\kappa}_w = 0.427$ (SE $= 0.0635$) with $\{w_{ij} = 1 - |i - j|/(c - 1)\}$ and $\hat{\kappa}_w = 0.458$ (SE $= 0.072$) with $\{w_{ij} = 1 - (i - j)^2/(c - 1)^2\}$. The difference between the weighted agreement and that expected under independence is less than half of the maximum possible. The agreement between Siskel and Ebert was not especially strong.

A disadvantage of kappa and weighted kappa is that their values depend strongly on the marginal distributions. The same diagnostic rating process can yield quite different values for κ or for κ_w, depending on the proportions of cases of the various types. Values of $\hat{\kappa}_w$ for different tables should be compared only if they use the same weights and have similar margins. Graham and Jackson (1993) noted that $\hat{\kappa}_w$ describes association more than agreement and can be large even when no observations fall on the main diagonal. See also Exercise 8.6. In summarizing a contingency table by a single number, the reduction in information can be severe. It is helpful to construct models providing more detailed investigation of the agreement and disagreement structure rather than to depend solely on a summary index.

8.5.6 Agreement Among Multiple Raters

With several raters, ordinary loglinear models for the joint cross-classification for all the raters are not usually relevant. With such models, the description of agreement and association between two raters is conditional on the ratings by the other raters. It is usually more relevant to study agreement marginally, without conditioning on

the other ratings. Then, simultaneously modeling the pairwise agreement structure requires modeling all the two-way marginal tables of the joint table for all the raters (Becker and Agresti 1992).

An alternative simple approach averages a pairwise measure of agreement over all the possible pairs of raters. Generalizations of kappa summarize pairwise agreements or multiple agreements (Landis and Koch 1977a,b; O'Connell and Dobson 1984). A related approach summarizes the concordance among the ratings, by using an adjustment of the Kendall *coefficient of concordance* to recognize ties and adjust for chance agreement (Fagot 1994). A more general approach models the value of a summary measure such as weighted kappa between pairs of raters as a function of explanatory variables, allowing that not all raters need observe all subjects (Gonin et al. 2000).

Latent variable models are also useful. One approach uses an ordinal categorical latent variable, with linear-by-linear association between the latent variable and each observed ordinal variable (Agresti and Lang 1993b). The number of latent classes can be a parameter, or it can be fixed. Another approach assumes continuous underlying evaluations which transform to ordered category ratings according to category thresholds for the rating categories. Such models separately evaluate the association between raters and the differences in their category definitions. See Uebersax (1993), Uebersax and Grove (1993), Qu et al. (1995), and Williamson and Manatunga (1997). Finally, the intraclass correlation can be estimated for underlying assumed continuous responses.[3]

8.6   MODELING ORDINAL PAIRED PREFERENCES

Sometimes categorical outcomes result from pairwise evaluations. A common example is athletic competitions, when the outcome for a team or player facing another consists of categories (win, lose) or (win, tie, lose). Another example is pairwise evaluation of product brands, such as brands of wine of some type. When a wine critic rates r brands of chianti classico, it might be difficult to establish an outright ranking, especially if r is large. However, after tasting any given pair of brands (say, a and b) at the same occasion, the critic could likely compare a to b on a scale such as (much worse, slightly worse, about the same, slightly better, much better). An overall ranking of the brands of wine could then be estimated using all of the pairwise preference evaluations. In this section we present a model for doing this. Below we refer to the items being compared (such as different brands of wine) using the generic term *treatments*.

8.6.1   Ordinal Extensions of the Bradley–Terry Model

Bradley and Terry (1952) proposed a logistic model for pairwise evaluations with a binary outcome. Of the r treatments, let Π_{ab} denote the probability that a is

---

[3]See john-uebersax.com/stat/agree.htm for a survey of agreement analyses.

preferred to b. Suppose that $\Pi_{ab} + \Pi_{ba} = 1$ for all pairs; that is, a tie cannot occur. The Bradley–Terry model is

$$\log \frac{\Pi_{ab}}{\Pi_{ba}} = \beta_a - \beta_b,$$

with a constraint such as $\beta_r = 0$. For this model, $\Pi_{ab} = \frac{1}{2}$ when $\beta_a = \beta_b$ and $\Pi_{ab} > \frac{1}{2}$ when $\beta_a > \beta_b$. The model reduces the set of $r(r-1)/2$ pairwise probabilities $\{\Pi_{ab}\}$ to the $r - 1$ treatment parameters $\{\beta_i\}$.

The Bradley–Terry model extends to comparisons with an ordinal scale. With cumulative logits and a c-category evaluation scale, let Y_{ab} denote the response for a comparison of a with b. The model, which can be motivated by an underlying latent variable model with logistic distributions (Tutz 1986), is

$$\text{logit } [P(Y_{ab} \le j)] = \alpha_j + (\beta_b - \beta_a), \tag{8.21}$$

with a constraint such as $\beta_r = 0$. If there is no effect relating to the order in which subjects observe treatments, $P(Y_{ab} \le j) = P(Y_{ba} > c - j) = 1 - P(Y_{ba} \le c - j)$. It follows that

$$\text{logit } [P(Y_{ab} \le j)] = -\text{logit } [P(Y_{ba} \le c - j)],$$

and necessarily $\alpha_j = -\alpha_{c-j}$. The most common ordered preference scale is (win, tie, lose). Then $\alpha_1 = -\alpha_2$.

If $\beta_a > \beta_b$, $P(Y_{ab} \le j) < P(Y_{ba} \le j)$. With an evaluation scale in comparing a and b such as (much worse, slightly worse, the same, slightly better, much better), this means that a is ranked more highly, in the sense that it is less likely that a is much worse than b than it is that a is much better than b. If $\beta_a = \beta_b$, a and b are equivalent in the sense that $P(Y_{ab} = j) = P(Y_{ab} = c - j + 1) = P(Y_{ba} = j)$ for all j.

This model can use an alternative link function, such as the probit. Another possibility applies the logit link to adjacent response probabilities (Agresti 1992a),

$$\log \frac{P(Y_{ab} = j)}{P(Y_{ab} = j + 1)} = \alpha_j + (\beta_b - \beta_a). \tag{8.22}$$

Again assuming no order effect, this logit is the same as $\log[P(Y_{ba} = c - j + 1)/P(Y_{ba} = c - j)]$, so that $\alpha_j = -\alpha_{c-j}$. The model satisfies

$$\frac{P(Y_{ab} = j)}{P(Y_{ba} = j)} = \exp[(c + 1 - 2j)(\beta_b - \beta_a)].$$

When $c = 5$, for example, suppose that $\lambda = P(Y_{ab} = 4)/P(Y_{ba} = 4) = \exp[2(\beta_a - \beta_b)]$ is the odds that a is considered slightly better than b instead of slightly worse. Then $\lambda^2 = P(Y_{ab} = 5)/P(Y_{ba} = 5)$ is the odds that a is much

better than b instead of much worse. The interpretation refers directly to an odds for a given outcome rather than an odds for two groupings of outcomes as occurs with cumulative logit models. This is more natural for many applications. The two forms of model share the property that Y_{ab} is stochastically higher than Y_{ba} when $\beta_a > \beta_b$.

Model (8.22) treats each pair of adjacent responses identically in terms of the effect comparing a and b. More generally, we could permit a positive distance d_j between categories j and $j + 1$, where $d_j = d_{c-j}$, for $j = 1, \ldots, c - 1$. A relatively small distance diminishes influences of treatment effects. This leads to a generalization of model (8.22) satisfying

$$\frac{P(Y_{ab} = j)}{P(Y_{ba} = j)} = \exp[(v_{c-j+1} - v_j)(\beta_b - \beta_a)], \tag{8.23}$$

where $\{v_j\}$ are monotone scores satisfying $\{v_{j+1} - v_j = d_j\}$.

8.6.2 Paired Preference Model Fitting and Inference

For a sample with pairwise evaluations of r treatments, let $n_{(ab)j}$ denote the number of times that outcome j occurs in comparing treatment a with treatment b, for each $a < b$ and $j = 1, \ldots, c$. Let $\{\mu_{(ab)j}\}$ denote expected frequences for the $r(r - 1)/2 \times c$ contingency table in which each row displays comparisons for a particular pair of treatments. To fit a paired preference model, the simplest approach treats each set $(n_{(ab)1}, n_{(ab)2}, \ldots, n_{(ab)c})$ with $a < b$ as an independent multinomial sample. The model can then be fitted with standard software, setting up the model matrix with the constraint $\alpha_j = -\alpha_{c-j}$ (see the Appendix for an example).

We can test goodness of fit by comparing observed counts $\{n_{(ab)j}\}$ with fitted values $\{\hat{\mu}_{(ab)j}\}$ for the model. The usual X^2 and G^2 fit statistics have asymptotic chi-squared distributions with df $= \binom{r}{2}(c - 1) - [(r - 1) + (c - 1)/2]$ when c is odd and df $= \binom{r}{2}(c - 1) - [(r - 1) + (c - 2)/2]$ when c is even. We can test the hypothesis of equivalent treatments by the change in the deviance between the null model with $\beta_1 = \cdots = \beta_r$ and the full model, using the chi-squared distribution with df $= r - 1$.

We elaborate further now on aspects of model fitting for the adjacent-categories logit model. In its general form (8.23), the model has the same fit as the loglinear model

$$\log \mu_{(ab)j} = \lambda + \lambda_{(ab)}^X + \lambda_j^Y + v_j(\beta_a - \beta_b),$$

where for all j, $\lambda_j^Y = \lambda_{c-j+1}^Y$. This loglinear model is a special case of the *row effects model* (6.5) presented in Section 6.3.1. Let

$$n_{(ab)+} = \sum_j n_{(ab)j}, \qquad n_{(+)j} = \sum_{a<b} \sum n_{(ab)j},$$

$$n_{(g)j} = \sum_{a>g} n_{(ga)j} + \sum_{a<g} n_{(ag)c-j+1}, \qquad n_{(g)+} = \sum_j n_{(g)j}.$$

Then $n_{(g)j}$ is the number of occurrences of outcome j in comparing treatment g with all other treatments. Suppose that we treat the $n_{(ab)+}$ comparisons of treatments a and b as independent multinomial trials and suppose that comparisons of different pairs of treatments are independent. Then cell counts in different rows of $\{n_{(ab)j}\}$ are independent multinomial samples, and the likelihood equations for model (8.23) are

$$\hat{\mu}_{(ab)+} = n_{(ab)+} \qquad \text{for all } a < b,$$

$$\hat{\mu}_{(+)j} + \hat{\mu}_{(+)c-j+1} = n_{(+)j} + n_{(+)c-j+1}, \qquad j = 1, \ldots, c,$$

$$\sum_j v_j \hat{\mu}_{(g)j} = \sum_j v_j n_{(g)j}, \qquad g = 1, \ldots, r.$$

For the scores $\{v_j\}$, the mean response when treatment g is compared with other treatments is the same for the observed and fitted data. For this model, the ML estimates $\{\hat{\beta}_g\}$ have the same order as these sample means.

The last set of likelihood equations results from differentiating the log likelihood with respect to $\{\beta_g\}$. Let

$$M_g = \sum_j v_j n_{(g)j}, \qquad g = 1, \ldots, r.$$

For the hypothesis of no treatment effects, H_0: $\beta_1 = \cdots = \beta_r$, let $\pi_1, \ldots, \pi_c$ with $\pi_j = \pi_{c-j+1}$ denote the response distribution for each pair of treatments. Without loss of generality, we let $\{v_j\}$ satisfy $v_j = -v_{c-j+1}$. Then, under H_0, $E(M_g) = 0$, $\text{Var}(M_g) = n_{(g)+} \sum_j v_j^2 \pi_j$, and $\text{Cov}(M_a, M_b) = -n_{(ab)+} \sum_j v_j^2 \pi_j$. When each pair of treatments has the same number of observations, the correlation for each (M_a, M_b) pair is $-1/(r-1)$, and the score statistic for testing H_0 has the simple form

$$S = \frac{(r-1) \sum_g M_g^2}{2 \sum_j v_j^2 n_{(+)j}}.$$

Its asymptotic null distribution is chi-squared with df $= r - 1$.

8.6.3 Example: Comparing Tastes of Soft Drinks

In 1985 the Coca-Cola Company introduced a sweeter formulation designed to replace its flagship soft drink, Coca-Cola (Coke). We refer to this newly formulated drink here as "New Coke" and the flagship drink as "Classic Coke." Table 8.8, from Agresti (1992a), refers to a soft-drink tasting experiment in which each of 61 subjects made three pairwise evaluations of A = New Coke, B = Classic Coke, and C = Pepsi. For comparing drink a with drink b, the experiment used the rating scale (much worse, slightly worse, about the same, slightly better, much better). In Table 8.8, the response sequence (1, 2, 5) for (BA, CA, CB), for instance, means that the subject rated Classic Coke much worse than New Coke, Pepsi worse than New Coke, and Pepsi much better than Classic Coke.

TABLE 8.8. Data for Soft-Drink Tasting Experiment[a]

BA	CA	CB	BA	CA	CB	BA	CA	CB	BA	CA	CB	BA	CA	CB
1	1	2	2	1	1	2	3	3	3	3	4	4	3	3
1	1	4	2	1	5	2	3	4	3	4	1	4	3	4
1	2	2	2	1	5	2	3	4	3	4	2	4	3	4
1	2	3	2	1	5	2	4	3	3	4	4	4	3	5
1	2	3	2	2	1	2	4	4	3	5	2	4	4	2
1	2	3	2	2	3	2	4	4	3	5	4	4	4	2
1	2	3	2	2	3	2	4	5	4	1	1	4	4	4
1	2	5	2	2	4	3	1	2	4	1	4	4	5	2
1	4	4	2	2	5	3	2	1	4	1	4	5	1	3
1	4	5	2	2	5	3	2	3	4	2	2	5	2	1
1	4	5	2	3	2	3	2	3	4	2	5	5	5	1
1	5	4	2	3	2	3	3	3	4	3	2	5	5	3
1	5	5												

Source: Agresti (1992a).
[a] A, New Coke; B, Classic Coke; C, Pepsi.

TABLE 8.9. One-Way Margins for Soft-Drink Tasting Evaluations from Table 8.8

	Preference Scale				
Pair (a, b)	a Much Worse	a Worse	No Preference	a Better	a Much Better
Classic Coke, New Coke	12	19	11	14	5
Pepsi, New Coke	11	18	12	13	7
Pepsi, Classic Coke	7	12	14	16	12

The joint ratings for the 61 subjects produce counts in a sparse contingency table having $5^3 = 125$ cells. Table 8.9 shows the one-way margins to which the model applies. For model fitting, we first treat each pairwise evaluation by a subject as independent of all others by the same subject or by other subjects. With this approach, we fit the model to Table 8.9 by treating the counts in that table as three independent multinomial samples.

We fitted the adjacent-categories logit model (8.22) by using software (see Table A.6 in the Appendix) to fit the equivalent baseline-category logit model

$$\log \frac{P(Y_{ab} = j)}{P(Y_{ab} = c)} = \alpha_j^* + (j - c)(\beta_a - \beta_j),$$

where α_j^* relates to α_j in (8.22) by

$$\alpha_j^* = \alpha_j + \alpha_{j+1} + \cdots + \alpha_{c-1}.$$

Because $\alpha_j = -\alpha_{c-j}$ for model (8.22), with $c = 5$ it follows that $\alpha_1^* = 0$ and $\alpha_2^* = \alpha_4^*$. Hence, the model can be expressed as

$$L = X\beta,$$

where L is the 12×1 vector of baseline-category logits

$$L = \left(\log \frac{P(Y_{12} = 1)}{P(Y_{12} = 5)}, \log \frac{P(Y_{12} = 2)}{P(Y_{12} = 5)}, \log \frac{P(Y_{12} = 3)}{P(Y_{12} = 5)}, \ldots, \log \frac{P(Y_{23} = 4)}{P(Y_{23} = 5)} \right)',$$

$\beta = (\alpha_2^*, \alpha_3^*, \beta_1, \beta_2)'$ (setting $\beta_3 = 0$), and the model matrix

$$X = \begin{pmatrix} 0 & 0 & -4 & 4 \\ 1 & 0 & -3 & 3 \\ 0 & 1 & -2 & 2 \\ 1 & 0 & -1 & 1 \\ 0 & 0 & -4 & 0 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & -2 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -4 \\ 1 & 0 & 0 & -3 \\ 0 & 1 & 0 & -2 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$$

The model fits well, with deviance $= 1.3$ (df $= 8$). The likelihood-ratio statistic for testing H_0: $\beta_1 = \beta_2 = \beta_3$ is 6.68 and the score statistic is $S = 6.86$ (df $= 3$, P-values $= 0.08$). Table 8.10 shows results of ML fitting.[4] This analysis ranks the drinks in the order (New Coke, Pepsi, Classic Coke) from most preferred to least preferred. However, the evidence of a difference is notable only between New Coke and Classic Coke. (A public uproar followed the announcement about the possible

TABLE 8.10. Comparisons from Fitting Adjacent-Categories Logit Model to Table 8.8, with SE Values in Parentheses and Results of Test of Homogeneity, H_0: $\beta_1 = \beta_2 = \beta_3$

Pair of Soft Drinks	Parameters	Independent ML	Dependent ML
New Coke, Classic Coke	$\hat{\beta}_1 - \hat{\beta}_2$	0.217 (0.084)	0.234 (0.094)
New Coke, Pepsi	$\hat{\beta}_1 - \hat{\beta}_3$	0.103 (0.082)	0.111 (0.079)
Classic Coke, Pepsi	$\hat{\beta}_2 - \hat{\beta}_3$	−0.113 (0.082)	−0.123 (0.090)
Test of homogeneity		6.68	6.39

[4]The results shown in Agresti (1992a) are incorrect, referring to a subset of the data originally obtained for only 53 of the subjects.

replacement of Classic Coke by New Coke, and the company soon returned to the classic formulation and eventually discontinued selling New Coke.)

8.6.4 Model Fitting Allowing Dependent Ratings

In some applications it is more sensible to treat different evaluations by the same subject (rater) as correlated rather than independent. When the separate evaluations by each subject are available, as in Table 8.8, we can investigate this dependence by fitting the models in a different manner. Let $T = \binom{r}{2}$ be the number of pairs of treatments. When all n subjects evaluate each pair of treatments with a c-category scale, the frequencies of the joint ratings can be displayed in a c^T contingency table. The models presented in this section apply to the one-dimensional margins of that table. When there is within-subject dependence of evaluations but we treat the one-way margins as independent samples, the ML estimates of model parameters are still consistent (assuming that the model holds), but the standard errors are inappropriate.

One approach to fitting the model while allowing dependence treats the cell counts in the c^T table as having a multinomial distribution and maximizes that multinomial likelihood subject to the constraint that the model holds for the marginal probabilities. This approach can use methodology developed by Lang and Agresti (1994), as discussed in Agresti (1992a) and available with the R function mph.fit discussed in the Appendix, as shown at www.stat.ufl.edu/~aa/ordinal/ord.html. When the joint evaluations table of size c^T is huge, a simpler approach uses the GEE methodology presented in Chapter 9, treating the T evaluations by a subject as a cluster of correlated observations.

For the soft-drink tasting data, Table 8.10 also shows results of ML fitting of the adjacent-categories logit model (8.22) by treating the samples as dependent. The estimated differences $\hat{\beta}_a - \hat{\beta}_b$ and their SE values are similar as in treating the three taste evaluations by each subject as independent observations. The model also fits well for this sampling model, with deviance = 2.49 (df = 8).

CHAPTER NOTES

Section 8.2: Models Comparing Matched Marginal Distributions

8.1. The marginal cumulative logit model (8.2) of proportional odds form, the corresponding subject-specific model (8.3), the ordinal quasi-symmetry model (8.7), and the conditional symmetry model (8.9) all imply stochastically ordered marginal distributions. Assuming merely a stochastic ordering of the marginal probabilities without a particular model form, Robertson et al. (1988, p. 290) discussed projections that provide ML fitted values. El Barmi and Dykstra (1995) tested for this stochastic ordering against the unrestricted alternative. Gao and Kuriki (2006) and Klingenberg et al. (2009) tested marginal homogeneity against stochastically ordered margins. Agresti and Coull (1998) described ways of testing marginal homogeneity against various order-restricted alternatives.

Section 8.3: Models for the Joint Distribution in a Square Table

8.2. Kateri and Agresti (2007) showed that given the marginal means, the ordinal quasi-symmetry model is the closest model to the symmetry model in terms of the Kullback–Leibler distance. Becker (1990b) and Agresti and Lang (1993b) presented other quasi-symmetric models. Hout et al. (1987) applied diagonals-parameter symmetry models. McCullagh (1978) and Goodman (1979b) also discussed the conditional symmetry model. Goodman (1985) used association models for the joint distribution in square tables. Sobel (1988) extended these models to stratified square tables and to $c \times c \times c$ tables. Plackett and Paul (1978) proposed Dirichlet latent structure models such that, conditionally on the category probabilities, the responses are independent, but unconditionally, when averaged with respect to the Dirichlet a positive trend occurs along the main diagonal. They incorporated ordinality by using a generalized Dirichlet mixture.

8.3. Specialized models apply to triangular tables that have observations only above or only below the main diagonal. Goodman (1968), Bishop and Fienberg (1969), and Altham (1975) showed how to fit a quasi-independence model. Sarkar (1989) and Tsai and Sen (1995) proposed tests of quasi-independence for such tables, using ordinal alternatives such as local log odds ratios that are uniformly of one sign. Goodman (1972) defined a generalized triangular model that satisfies quasi-independence above the main diagonal and a separate quasi-independence below the main diagonal.

Section 8.4: Comparing Marginal Distributions for Matched Sets

8.4. The ordinal quasi-symmetry models (8.15) and (8.18) were proposed by Agresti (1993a), with the logistic linear trend version (8.7) for matched pairs in Agresti (1983c). Agresti (1993b) and Agresti and Lang (1993a) considered analogous models with cumulative logits. Klotz (1980) extended the *Cochran Q* statistic for comparing several matched proportions to the case of ordered response categories with some missing observations. His test generalizes the *Friedman test* to accommodate tied observations and missingness.

Section 8.5: Analyzing Rater Agreement on an Ordinal Scale

8.5. The following surveyed measuring and modeling agreement with ordinal data: Banerjee et al. (1999) presented summary measures, Roberts and McNamee (2005) focused on kappa-type measures, Schuster and von Eye (2001) focused on models, and von Eye and Mun (2005) presented a chapter on the loglinear modeling approach. Broemeling (2009) presented Bayesian approaches. Perkins and Becker (2002) simultaneously modeled univariate marginal responses and bivariate marginal associations to evaluate agreement both in terms of the overall frequency of responses and the category-specific agreement among pairs of raters. Ordinal association models have been used to summarize agreement (Becker and Agresti 1992; Valet et al. 2007) and to study social and occupational mobility (Duncan 1979; Goodman 1979c; DiPrete 1990; Xie 1992, Lang and Eliason 1997, Sobel et al.

1998). For Bayesian approaches, see Johnson (1996), Mwalili et al. (2004), and Kottas et al. (2005).

Section 8.6: Modeling Ordinal Paired Preferences

8.6. Böckenholt and Dillon (1997) presented an alternative way to handle dependence with ordinal paired preference evaluations. They proposed a latent class extension of adjacent-categories logit and cumulative probit models that accounts for preference differences among the raters. Dittrich et al. (2004) noted that preference decisions typically depend on characteristics of the raters and the subjects being rated. They extended adjacent-categories logit paired preference models in loglinear form by incorporating categorical rater- and subject-specific information. Dittrich et al. (2007) extended this, also incorporating dependencies and applying the paired preference structure to make comparisons of a set of Likert responses on a common scale. Fahrmeir and Tutz (1994) modeled paired comparisons over time, allowing effects to change with time.

EXERCISES

8.1. Suppose that (Y_1, Y_2) has bivariate normal probability density function $f(y_1, y_2)$, with $E(Y_1) = \mu$, $E(Y_2) = \mu + \Delta$, $\text{Var}(Y_1) = \text{Var}(Y_2) = \sigma^2$, and $\text{Corr}(Y_1, Y_2) = \rho$. Show that $f(y_1, y_2)/f(y_2, y_1)$ has form $\delta^{y_1 - y_2}$ for some constant δ. Thus, under the assumption of an underlying bivariate normal distribution, explain why the ordinal quasi-symmetry model (8.7) may be appropriate for a square ordinal table.

8.2. For matched pairs, consider the special case (8.18) of the loglinear ordinal quasi-symmetry model (8.15), namely

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \beta_1 u_i + \beta_2 u_j + \rho_{i+j}.$$

For equally spaced scores, explain why this model has the form

$$\mu_{ij} = \alpha_i \alpha_j \delta_{i-j} \delta^*_{i+j},$$

which is in a class of models proposed by Goodman (1985).

8.3. Consider the conditional symmetry (CS) model (8.9).

(a) Show that this model has the loglinear representation

$$\log \mu_{ij} = \lambda_{\min(i,j),\max(i,j)} + \beta I(i < j),$$

where $I(\cdot)$ is an indicator (see also Bishop et al. 1975, pp. 285–286).

(b) Show that conditional symmetry + marginal homogeneity = symmetry. Explain why $G^2(S \mid CS) = G^2(S) - G^2(CS)$ tests marginal homogeneity (df = 1). When the model holds, $G^2(S \mid CS)$ is more powerful asymptotically than $G^2(S \mid QS)$. Why?

8.4. Formulate a marginal model using adjacent-categories logits or continuation-ratio logits that is analogous to cumulative logit model (8.11). Interpret parameters.

8.5. Table 8.11, from a general social survey, was analyzed with the OQS model by Agresti (1995). Subjects gave their opinions regarding government spending on (1) the environment, (2) health, (3) assistance to big cities, and (4) law enforcement. Analyze the data using an alternative model such as a marginal cumulative logit or adjacent-categories logit model.

TABLE 8.11. Opinions About Government Spending^a

Cities:		1			2			3		
Law Enforcement:		1	2	3	1	2	3	1	2	3
Environment	Health									
1	1	62	17	5	90	42	3	74	31	11
	2	11	7	0	22	18	1	19	14	3
	3	2	3	1	2	0	1	1	3	1
2	1	11	3	0	21	13	2	20	8	3
	2	1	4	0	6	9	0	6	5	2
	3	1	0	1	2	1	1	4	3	1
3	1	3	0	0	2	1	0	9	2	1
	2	1	0	0	2	1	0	4	2	0
	3	1	0	0	0	0	0	1	2	3

Source: General Social Survey.
^{a} 1, too little; 2, about right; 3, too much.

8.6. For weighted kappa with $\{w_{ij} = 1 - (i - j)^2/(c - 1)^2\}$, show that two tables that have the same marginal distributions have the same values of κ_w whenever they have the same correlations, with row and column numbers as the scores (Graham and Jackson 1993).

CHAPTER 9

Clustered Ordinal Responses: Marginal Models

Many studies observe a response variable for clusters of subjects. The matched-pairs type of observation analyzed in Chapter 8 is the simplest type of clustered data, with each pair forming a cluster. In practice, there may be more than two observations in a cluster and there are usually also explanatory variables.

Longitudinal studies are a common source of clustered data, with the repeated observations on a subject at different times forming a cluster. For example, a physician might evaluate patients who are using a placebo or a new drug treatment for a curable condition at weekly intervals using the scale (cured, improved, no change, worse). The set of responses for a particular subject forms a cluster. In this chapter we present examples with repeated observations of an ordinal response from longitudinal clinical trials to analyze progress over time in Section 9.2.3 for treating arthritis and in Section 9.3.3 for treating insomnia. Repeated responses on a subject need not refer to different times, however. A dental study might make a visual observation of the extent of plaque (none, mild, moderate, severe) for each tooth in a subject's mouth. Analyses should take the clustering into account, as two teeth in the same person's mouth are likely to be more similar than two teeth from different mouths.

In some applications groupings of like subjects form clusters. For example, a study of factors that affect children's weight, measured with the ordinal scale (normal, overweight, obese), might sample families and treat children from the same family as a cluster. Again, observations within a cluster tend to be more similar than observations from different clusters. Inferential analyses that ignore the clustering may be badly biased.

In most studies with clustered data, besides comparing the marginal distributions as we did in Chapter 8, it is also of interest to analyze effects of explanatory variables on the marginal responses. For instance, a longitudinal study might analyze

---

whether responses tend to improve over time and whether different age or socioeconomic subgroups have different trends. In Chapters 10 and 11 we show how to incorporate the effects of explanatory variables. The main focus of this chapter is on *marginal modeling*, emphasizing the *generalized estimating equations* (GEE) approach for parameter estimation, which is computationally simpler than ML. We also consider a *transitional* approach for longitudinal studies that models observations at a given time by including previous outcomes among the explanatory variables. Then in Chapter 10 we present an alternative approach using cluster-level random effect terms in the model. Such models have *conditional*, *cluster-specific* interpretations that apply at the cluster level. This contrasts with marginal models, which have *population-averaged* interpretations, a distinction introduced in Section 8.2.3.

As in Section 8.4, we let T denote the number of observations in a cluster. The clusters may have different sizes. Sometimes this happens because of the nature of the cluster, such as when each cluster is a family. Even when a study designs the clusters to have the same size, in practice the observed data often have clusters of different sizes because data are missing for some observations in some clusters. For simplicity of notation in this chapter, however, we express the cluster size for cluster i as T rather than T_i.

9.1 MARGINAL ORDINAL MODELING WITH EXPLANATORY VARIABLES

At observation Y_t in $Y_1, Y_2, \ldots, Y_T$ on a c-category scale, we can model the marginal response distribution using $c - 1$ cumulative logits or probits or some other type of ordinal link function. For a particular link function, a model for Y_t has the form

$$\text{link}_{jt} = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}_t, \qquad j = 1, \ldots, c - 1, \quad t = 1, \ldots, T,$$

such as with $\text{link}_{jt} = \text{logit } [P(Y_t \le j)]$ or $\log[P(Y_t = j)/P(Y_t = j + 1)]$. When we replace $\boldsymbol{\beta}_j$ by $\boldsymbol{\beta}$, the model takes the proportional odds form with the same effects for each logit. With cumulative logit link, this is the marginal model of the form

$$\text{logit } [P(Y_t \le j)] = \alpha_j + \boldsymbol{\beta}' \mathbf{x}_t = \alpha_j + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}. \qquad (9.1)$$

Some parameters in $\boldsymbol{\beta}$ may refer to the variable subscripted by t (e.g., time) that indexes the clustered observations. One can then compare marginal distributions at particular settings of $\mathbf{x}$ or evaluate effects of $\mathbf{x}$ on the response. The linear predictor can also contain interaction terms: for example, to analyze whether the effects of $\mathbf{x}$ are the same at each t.

9.1.1 Within-Cluster and Between-Cluster Standard Errors

Ignoring the clustered nature of the data and, instead, treating the observations as independent results in invalid standard errors for the model parameter estimates. Whether a standard error is too large or too small typically depends on whether the parameter refers to a within-cluster effect or a between-cluster effect. For within-cluster effects, the same clusters occur in each of the groups compared. For between-cluster effects, the groups to be compared have distinct clusters. For example, consider the ordinal response Y = relief of headache (none, some, complete) measured in a crossover study for two drugs and for female and male subjects. Each subject forms a cluster that is evaluated for both drugs. A comparison of the two drugs on Y is a within-subject comparison, because in a crossover study each subject takes both drugs. A comparison of females and males on Y is a between-subject comparison, because female subjects are distinct from male subjects.

A within-cluster correlation is usually positive. For example, subjects who respond relatively well under one condition usually tend to respond relatively well in another. With positive within-cluster correlation, the true standard errors for within-cluster effects tend to be smaller than if the same effects occurred with the same number of independent observations. As in other settings, such as comparing two means with matched-pairs data, blocking the data into relatively homogeneous clusters results in improved precision of within-cluster effects.

By contrast, standard errors for between-cluster effects tend to be larger than if the same effects occurred with the same number of independent observations. Since T observations from within a cluster tend to be more alike than T observations from different clusters, those observations provide less information about a group than is provided by T independent observations. In Section 9.2.3 we show an example of SE bias in ignoring the clustering of data.

9.1.2 ML Fitting of Marginal Ordinal Models

As explained in Section 8.2.1, ML fitting of marginal models for categorical responses is not simple, because the models refer to marginal probabilities, whereas the likelihood function specifies the joint distribution of the clustered responses. So it is not usually possible to express the likelihood function in terms of the model parameters of interest. ML fitting becomes even more difficult when there are also explanatory variables.

At each combination of values of explanatory variables, we assume a multinomial distribution with c^T cell probabilities π to describe the distribution of the T observations on the c-category response. Marginal models that use ordinal logits for the margins can be expressed as special cases of the generalized loglinear model introduced in (6.18): namely,

$$C \log A\pi = X\beta.$$

The matrix A is a *marginalization matrix* that, when multiplied by π, forms the marginal probabilities to be modeled. Each row of A contains 1 and 0 elements in the appropriate positions to form the relevant marginal probabilities. For marginal

cumulative logit models, for example, this forms marginal cumulative probabilities and their complements. The matrix C is a *contrast matrix* that constructs the relevant logits. Each row of C has a 1 in one position, a -1 in one position, and 0 elsewhere, to form the appropriate contrast of logs of the marginal probabilities or their sums. This generalized loglinear model maps the c^T cell probabilities to a much smaller number of marginal logits, and it is not invertible. That is, we cannot express the multinomial probabilities π in terms of the marginal model parameters β, so standard methods are not available for maximizing that multinomial likelihood function.

Molenberghs and Lesaffre (1999) and Agresti (2002, pp. 464–466) surveyed methods for ML fitting of marginal models. One approach views the model as a constraint equation and uses methods for maximizing a likelihood function subject to constraints. The method introduces Lagrange multipliers corresponding to the constraints and solves the Lagrangian likelihood equations using a Newton–Raphson algorithm (Lang and Agresti 1994, Lang 1996, 2004, 2005). Another approach uses a one-to-one correspondence between joint cell probabilities and parameters that describe the marginal distributions, the bivariate distributions, the trivariate distributions, and so on (Molenberghs and Lesaffre 1994). Multivariate logistic models apply then to the component distributions, although some higher-order effects may be assumed to vanish, for simplicity.

Compared to the continuous-response case using the multivariate normal, the number of parameters can be extremely large. For example, with T means and a common variance and correlation, the multivariate normal has only $T + 2$ parameters (ignoring, for now, the parameters for modeling the explanatory variables). By contrast, in the categorical case, there are $c^T - 1$ parameters for the multinomial at each setting of the explanatory variables, unless the joint distribution is itself modeled in addition to the marginal distribution. ML fitting is not practical when T is large or when there are many predictors, especially when some are continuous. It is even more difficult to implement ML when the number of observations is not the same in each cluster. Currently, ML is available only in certain cases with specialized software, such as the R function mph.fit described in the Appendix.

One way in which it is more feasible to conduct ML fitting of marginal models is to specify a latent variable model such that the likelihood function can be expressed directly in terms of model parameters. For example, Kim (1995) developed such models for applications in opthalmology with bivariate ordinal categorical responses. His bivariate probit model assumed an underlying bivariate normal distribution for the two ordinal responses and specified an ordinary linear model for each latent normal response in terms of the explanatory variables.

Williamson and Kim (1996) proposed similar models without assuming normality for the latent variables, instead describing association by global odds ratios. They argued that such an approach could be more flexible, as the joint distribution need not be normal and various models can apply to the marginal probabilities, including cumulative logit and probit models. For the models they presented, it is possible to express the likelihood function in terms of the model parameters of interest, so ML fitting is not as difficult.

9.1.3 Example: Eye Disease Risk Factors

Williamson and Kim (1996) analyzed data from a Wisconsin epidemiological study
in which insulin-taking young diabetics were examined to assess the prevalence
and severity of diabetic retinopathy, which is a type of noninflammatory damage to
the retina of the eye. Table 9.1 shows the data for the right and left eyes, stratified
by the gender of the person, with the four ordered categories used to summarize
retinopathy severity. Several explanatory variables were measured in the study.
Some of them, such as gender and diastolic blood pressure, were characteristics of
the person. Others, such as a binary variable indicating whether an eye has macular
edema, were characteristics of the right or left eye and were observed for each eye.
Each person forms a cluster, with an observation on the response variable for each
eye. The complete data set is available at www.stat.ufl.edu/~aa/ordinal/ord.html.

Let Y_L and Y_R denote the retinopathy responses for the left and right eyes.
Let $\mathbf{x}_P$ denote person-specific explanatory variables, and let $\mathbf{x}_L$ and $\mathbf{x}_R$ denote
characteristics of the left and right eyes. For eye characteristics for which there is
no reason to expect a difference between right and left eyes, we could consider a
marginal model such as the cumulative logit model of proportional odds form,

$$\text{logit}\,[P(Y_L \le j)] = \alpha_j + \boldsymbol{\beta}_1' \mathbf{x}_P + \boldsymbol{\beta}_2' \mathbf{x}_L, \;\; \text{logit}\,[P(Y_R \le j)] = \alpha_j + \boldsymbol{\beta}_1' \mathbf{x}_P + \boldsymbol{\beta}_2' \mathbf{x}_R.$$

Because the eye-specific effects $\boldsymbol{\beta}_2$ are assumed to be the same for each eye, at
each setting of the explanatory variables the model implies marginal homogeneity
between the right and left eye margins.

Although the primary focus is on the effects of the explanatory variables on Y_L
and Y_R, we can also model the association between Y_L and Y_R and describe how it
depends on explanatory variables $\mathbf{x}$. Those explanatory variables may be a subset
of those in $\mathbf{x}_P$, $\mathbf{x}_L$, and $\mathbf{x}_R$. Let $\theta_{hj}^G(\mathbf{x})$ denote the global odds ratio at setting $\mathbf{x}$
of the explanatory variables for the binary collapsing of (Y_L, Y_R) following row h

TABLE 9.1. Severity of Diabetic Retinopathy, by Eye and Gender

| | | Retinopathy Severity | | | |
| | | Left Eye | | | |
Treatment	Right Eye	None	Mild	Moderate	Proliferative
Females	None	109	22	1	0
	Mild	16	108	21	3
	Moderate	0	20	30	4
	Proliferative	0	1	7	15
Males	None	128	15	0	0
	Mild	15	92	14	1
	Moderate	0	19	50	5
	Proliferative	0	0	4	20

Source: Williamson and Kim (1996).

and column j. A model that assumes a uniform global odds ratio at each setting of the explanatory variables has the form

$$\log \theta^G_{hj}(\mathbf{x}) = \alpha + \boldsymbol{\beta}'\mathbf{x},$$

in which the linear predictor does not depend on h or j. More generally, in allowing the global odds ratio to depend on these cutpoints for the collapsing, we could model the global odds ratios as being the same for cutpoints (h, j) in (Y_L, Y_R) as for cutpoints (j, h).

Some of the explanatory variables in this study are continuous, so a contingency-table representation of the data has a separate 4×4 table for each person, with a 1 in one cell and 0 in each other cell. The likelihood function is the product of these multinomial mass functions, each with a single observation. Each particular cell probability $\pi_{hj}(\mathbf{x}_i)$ for the 4×4 table for person i relates to the joint distribution function $F_{hj} = P(Y_L \le h, Y_R \le j)$ by

$$\pi_{hj}(\mathbf{x}_i) = F_{hj}(\mathbf{x}_i) - F_{h-1,j}(\mathbf{x}_i) - F_{h,j-1}(\mathbf{x}_i) + F_{h-1,j-1}(\mathbf{x}_i).$$

The joint distribution function can itself be expressed in terms of the global odds ratios and the marginal distribution functions using equation (2.8). Then, substituting in that expression the global odds ratio model for the global odds ratios and the cumulative logit models for the marginal distribution functions yields the product multinomial likelihood function in terms of the parameters $(\{\alpha_j\}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ for the marginal model and the parameters $(\alpha, \boldsymbol{\beta})$ for the association model. Williamson and Kim employed a quasi-Newton algorithm to maximize that function and obtain the ML estimates and their SE values.

Their model selection resulted in a model with estimates summarized in Table 9.2. The only explanatory variable needed for the global odds ratio association model was the person's gender. The estimated uniform global odds ratio for the nine possible collapsings of the 4×4 table for each person was $e^{3.61} = 37$ for males and $e^{3.61-0.81} = 16$ for females. The parameter estimates for the marginal model suggest that longer duration of diabetes, higher glycosylated hemoglobin level, higher diastolic blood pressure, presence of proteinuria, presence of macular edema, and male gender are jointly associated with worse retinopathy. The estimates have the usual interpretation for cumulative logit models. For example, given the other explanatory variables, the estimated odds that the retinopathy severity is less than any particular fixed level for females equals $e^{0.32} = 1.38$ times the corresponding estimated odds for males.

Williamson and Kim (1996) noted the challenge of checking the adequacy of the model, since it contains many continuous explanatory variables. One standard method that is available with ML model fitting is likelihood-ratio tests comparing the model to more general models with additional predictors such as interaction terms.

TABLE 9.2. ML Estimates from Fitting Model for Eye Data Using Cumulative Logits for Margins and Global Odds Ratio for Association

Effect	Estimate	SE
Model for margin		
Duration of diabetes (years)	−0.124	0.008
Glycosylated hemoglobin level	−0.093	0.026
Diastolic blood pressure	−0.041	0.007
Gender (1 = female, 0 = male)	0.320	0.140
Proteinuria (1 = present, 0 = absent)	−0.869	0.208
Macular edema (1 = present, 0 = absent)	−1.33	0.229
Model for association		
Intercept	3.61	0.180
Gender (1 = female, 0 = male)	−0.814	0.353

Source: Williamson and Kim (1996).

9.2 MARGINAL ORDINAL MODELING: GEE METHODS

For marginal models, ML fitting is computationally awkward and difficult for many data sets. An alternative to ML fitting uses a multivariate generalization of *quasi-likelihood*. For a univariate response, rather than assuming a particular probability distribution for Y, the quasi-likelihood method specifies only a linear model for a link function applied to $\mu = E(Y)$ and a formula $v(\mu)$ for how the variance of Y depends on the mean. The multivariate generalization does this for each Y_t for the marginal model and uses a guess for the correlation structure among $(Y_1, Y_2, \ldots, Y_T)$ without assuming a particular joint probability distribution.

9.2.1 Generalized Estimating Equation Methodology: Basic Ideas

For a multivariate response $Y_1, Y_2, \ldots, Y_T$, we first describe briefly the basic ideas of the quasi-likelihood method for marginal models in the case that each Y_t is one-dimensional, such as a binomial or Poisson variate. Then we describe the extension to an ordinal multinomial response.

For the marginal model for Y_t, the nature of Y_t usually suggests a particular variance function v. For example, when Y_t is a binary variable for a marginal logistic regression model, then $v(\mu_t) = \mu_t(1 - \mu_t)$. Various structures are possible for the *working correlation* matrix for the responses $Y_{i1}, Y_{i2}, \ldots, Y_{iT}$ in cluster i. A popular one is an *exchangeable* structure, which assumes a common value for Corr(Y_{is}, Y_{it}) for each pair (s, t) of observations in the cluster. That common correlation value is estimated from the data. In choosing a correlation structure, we attempt to capture the main component of the dependence in whatever joint distribution actually exists. Unlike in the multivariate normal case, specifying marginal means and variances and the correlation structure does not fully determine the multivariate distribution for the joint distribution of $Y_1, Y_2, \ldots, Y_T$.

The estimates of model parameters with this approach are solutions of equations that, in the way they incorporate the mean and variance structure, resemble likelihood equations. The equations are called *generalized estimating equations*, and the method is referred to as the *GEE method*. The equations are not likelihood equations, however, because the method does not specify a full multivariate distribution for $Y_1, Y_2, \ldots, Y_T$, so the method does not have a likelihood function.

When the marginal models hold, the GEE estimates of model parameters are valid even if the working correlation structure is misspecified. The validity of the GEE estimates holds even under the simplistic working correlation structure by which the observations are pairwise uncorrelated, that is, $\text{Corr}(Y_{is}, Y_{it}) = 0$ for each (s, t), referred to as *independence working correlations*. The estimates then equal those obtained using ML and treating different observations within a cluster as if they are independent observations, hence the same as different observations from different clusters. For binary data and count data, estimators based on this naive working correlation structure can have surprisingly good efficiency when the actual correlation is not very strong.

Although the model parameter estimates are valid, standard errors based solely on the assumed marginal and joint structure may not be, especially if the actual correlation structure is quite different from the working guess. For example, standard errors based on independence working correlations are badly biased if the responses are actually strongly associated. More appropriate standard errors result from an adjustment the GEE method makes using the empirical dependence exhibited by the data. The standard errors based on the working correlation structure are updated using this empirical dependence to provide more appropriate *robust* standard errors. These use a *sandwich covariance matrix* that is based on a product of three matrices, the ends of which are the covariance matrix if the working correlation structure were truly correct and the middle of which uses the empirical evidence. Unless the number of clusters is quite large, however, these empirically based standard errors can themselves tend to underestimate the true standard errors and have potentially large variability.

The GEE method is appealing for categorical data because of its computational simplicity compared to ML. However, the method has limitations. Since it does not have a likelihood function, likelihood-based methods such as likelihood-ratio tests and profile likelihood confidence intervals are not available for checking fit, comparing models, and conducting inference about parameters. Instead, inference uses Wald statistics. For example, the test of H_0: $\beta_k = 0$ treats the GEE estimate $\tilde{\beta}_k$ of β_k as having an approximate normal sampling distribution. Using a standard error SE from the sandwich-estimated covariance matrix, we obtain a P-value by referring $z = \tilde{\beta}_k / \text{SE}$ to a standard normal distribution or z^2 to a chi-squared distribution with df $= 1$. Some software (such as PROC GENMOD in SAS) also reports analogs of score tests for effects of predictors. These are more trustworthy than Wald tests.

9.2.2 GEE Approach: Ordinal Responses

The GEE methodology, originally introduced for modeling univariate marginal distributions such as the binomial, has since been extended to marginal modeling of multinomial responses. Lipsitz et al. (1994) proposed a GEE approach for cumulative logit models with repeated ordinal responses. We next outline this approach.

Marginally, we assume a multinomial distribution for each Y_t, $t = 1, \ldots, T$. For cluster i, each observation Y_{it} is specified by a set of $c - 1$ indicator variables, the jth one indicating whether Y_{it} falls in category j, $j = 1, \ldots, c - 1$. Specifically, let $y_{ijt} = 1$ if observation t in cluster i has outcome j ($j = 1, \ldots, c - 1$), with $y_{ijt} = 0$ otherwise. Let $\mathbf{y}_i$ be the $T(c - 1)$ binary indicators for cluster i for the T observations. The covariance matrix $\mathbf{V}_i$ for $\mathbf{y}_i$ is a $T(c - 1) \times T(c - 1)$ matrix. The covariance matrix $\mathbf{V}_{it}$ for the $c - 1$ indicators for each Y_{it} is a $(c - 1) \times (c - 1)$ matrix block on the main diagonal of $\mathbf{V}_i$ that is the multinomial covariance matrix for a single multinomial trial. That is, the covariance matrix $\mathbf{V}_{it}$ for $y_{i1t}, \ldots, y_{i,c-1,t}$ has entry $v_{ijt} = P(Y_{ijt} = 1)[1 - P(Y_{ijt} = 1)]$ for the cell on the main diagonal in row j and column j and entry $-P(Y_{iht} = 1)P(Y_{ijt} = 1)$ for the cell in row h and column j with $h \neq j$.

The remaining elements of $\mathbf{V}_i$ contain elements $\mathrm{Cov}(Y_{iht}, Y_{ijs})$ for $s \neq t$ that are not determined by the marginal multinomial covariances. For each pair (h, j) of outcome categories, the GEE approach uses a working correlation matrix for the pairs (Y_{is}, Y_{it}) of clustered observations. The working covariance matrix $\mathbf{V}_i$ for $\mathbf{y}_i$ specifies a pattern for $\mathrm{Corr}(Y_{ijt}, Y_{ihs})$ for each pair of outcome categories (h, j) and each pair (s, t) of observations in a cluster. An awkward aspect is the large number of parameters in the working correlation structure, especially if c and/or T are large. One way to reduce this somewhat uses the exchangeable structure, whereby $\mathrm{Corr}(Y_{iht}, Y_{ijs}) = \rho_{hj}$ for all pairs (s, t) in a cluster. That is, for a given pair of categories, the correlation is the same for all pairs of observations in a cluster. This is still a substantial number of parameters if c is large. The independence working correlation structure is $\rho_{hj} = 0$ for all h and j.

Let $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$. This is a function of the model parameters $\boldsymbol{\beta}$ that depends on the choice of model. The generalized estimating equations for $\boldsymbol{\beta}$ are

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_i \right) = \mathbf{0},$$

where $\mathbf{D}_i' = \partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}$. Lipsitz et al. (1994) suggested a Fisher scoring algorithm for solving these equations and a method of moments update for estimating $\{\rho_{hj}\}$ at each step of the iteration.

As in the univariate case, the GEE method uses the empirical dependence to find sandwich-covariance-based standard errors that are appropriate with large samples even if the working correlation guess is poor. For example, standard errors based on assuming independent observations would usually be invalid. An empirically

adjusted sandwich covariance matrix for the GEE estimate $\hat{\beta}$ is

$$\left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathrm{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i\right] \left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1}.$$

This is itself estimated by substituting $\hat{\mu}_i$ from the model fit in $\mathbf{D}_i$ and $\mathbf{V}_i$ and replacing $\mathrm{Cov}(\mathbf{Y}_i)$ by the empirical covariance matrix of $\mathbf{y}_i$.

9.2.3   Example: Arthritis Clinical Trial

To illustrate ordinal GSS methods, Lipsitz et al. (1994) used data from a randomized clinical trial comparing a drug (auranofin) with placebo for the treatment of rheumatoid arthritis. Patients were randomly assigned to a treatment, which they received throughout the study. At the start of the study and after one month, three months, and five months of treatment, patients assessed their arthritis on the ordinal scale (good, fair, poor). Other explanatory variables measured were age and gender. Table 9.3 shows the data for four of the 303 patients in the study. The complete data set is available at www.stat.ufl.edu/~aa/ordinal/ord.html.

Lipsitz et al. treated the baseline self-assessment as a covariate and modeled the three follow-up assessments as a function of that baseline assessment, treatment, age, sex, and time of observation. Let Y_t denote follow-up response t ($t = 1, 2, 3$). Each subject forms a cluster, with the observations (y_{i1}, y_{i2}, y_{i3}) in cluster i being the ordinal response at the three follow-up occasions of observation. Table 9.4 shows the sample marginal distributions of the ordinal arthritis assessment at the three follow-up times and at the baseline. At the baseline, the distributions are similar for the two treatments, as expected by the randomization of patients to treatments. At follow-up times, the responses tend to improve, more for auronofin than for the placebo. The change does not seem to be uniformly monotone for the poor or the good outcomes, which suggests using indicator variables for the time effects rather than a linear trend. Relevant questions then include whether the response is

TABLE 9.3. Four Observations from Clinical Trial Comparing Arthritis Assessment for Two Treatments

| Subject | Sex | Age | Treatment | Arthritis Assessment[a] | | | |
				Baseline	1 Month	2 Months	3 Months
1	M	55	Auranofin	2	1	1	1
2	F	60	Auranofin	3	2	2	2
3	M	28	Placebo	2	3	1	3
4	F	47	Placebo	3	1	3	2

Source: Lipsitz et al. (1994). Complete data are at www.stat.ufl.edu/~aa/ordinal/ord.html.
[a] 1, good; 2, fair; 3, poor.

TABLE 9.4. Sample Marginal Distributions of Arthritis Assessment

	Auranofin Response (%)			Placebo Response (%)		
Occasion	Good	Fair	Poor	Good	Fair	Poor
Baseline	22	45	33	22	47	31
1 month	37	51	12	36	34	30
3 months	45	35	20	30	43	28
5 months	50	35	15	39	35	25

significantly better for auronofin than for the placebo and whether the treatment effect changes with time (i.e., whether there is treatment × time interaction).

Lipsitz et al. fitted the cumulative logit model of proportional odds form

$$\text{logit}\,[P(Y_t \le j)] = \alpha_j + \beta_1 t_2 + \beta_2 t_3 + \beta_3 b + \beta_4 a + \beta_5 s + \beta_6 tr,$$

with indicator variables t_2 for time 2 and t_3 for time 3 (with $t_2 = t_3 = 0$ for time 1), $b =$ baseline response (scored 1, 2, 3), $a =$ age, s for sex (1 = male, 0 = female), and tr for treatment (1 = auranofin, 0 = placebo). They noted that more complex models having interaction terms did not fit better. They showed that similar results occur with various working correlation structures. With the independence structure, the GEE approach gives the prediction equation

$$\text{logit}\,[\hat{P}(Y_t \le j)] = \hat{\alpha}_j - 0.046 t_2 + 0.266 t_3 - 1.035 b - 0.006 a + 0.133 s + 0.548 tr$$

with SE values (0.124, 0.118, 0.137, 0.008, 0.184, 0.176) for the explanatory effects. The important effects are for baseline and for treatment. A simpler model that smooths the effects by deleting those for time 2, age, and sex has the fit

$$\text{logit}\,[\hat{P}(Y_t \le j)] = \hat{\alpha}_j + 0.289 t_3 - 1.042 b + 0.554 tr,$$

with SE values 0.098, 0.136, and 0.176. Controlling for other variables, patients using auranofin have estimated odds of response good instead of fair or poor and estimated odds of response good or fair instead of poor that are exp(0.554) = 1.74 times the corresponding estimated odds for patients using the placebo. A change in the baseline response from 1 = good to 2 = fair or from 2 = fair to 3 = poor has the effect of multiplying the estimated odds of response in category j or below by exp(−1.042) = 0.35.

As explained in Section 9.2.2, the working correlation structure comprises the correlations of binary indicators for each pair of outcome categories between 1 and $c − 1$ for all pairs of observations (t, s) in a cluster. For example, the exchangeable structure has Corr$(Y_{iht}, Y_{ijs}) = \rho_{hj}$ for all t and s. For these data, the indicator variables for the first two categories at each of three times are y_{i11}, y_{i21}, y_{i12}, y_{i22}, y_{i13}, and y_{i23}. For all three pairs of times $(t, s) = (1,2)$, $(1,3)$, $(2,3)$, the estimated

exchangeable correlations for the pairs with the first two outcome categories are $\text{Corr}(Y_{i1t}, Y_{i1s}) = 0.39$, $\text{Corr}(Y_{i1t}, Y_{i2s}) = 0.15$, and $\text{Corr}(Y_{i2t}, Y_{i2s}) = 0.21$.

Recall from Section 9.1.1 that when there is positive within-cluster correlation, standard errors for within-cluster effects tend to be smaller, and standard errors for between-cluster effects tend to be larger than if the same effects occurred with the same number of independent observations. If we had naively treated repeated responses as independent for the entire analysis, the SE values for within-subject time effects would be misleadingly large and the SE values for between-subject effects would be misleadingly small. For example, the treatment effect estimate would have had SE = 0.130 rather than 0.176.

For simpler interpretation, it can be helpful to report sample marginal means and their differences. With arthritis assessment scores (1, 2, 3), the initial means were 2.10 for auranofin and 2.09 for the placebo. The means after five months were 1.65 for auranofin and 1.86 for the placebo. The difference in means between the baseline and five-month responses was 0.45 for auranofin and 0.23 for the placebo.

9.2.4 Modeling Association to Generate Working Correlations

The GEE approach requires selecting a working correlation matrix for the binary category indicators. Disadvantages are that this can involve a large number of parameters and that the correlation is not a natural measure for binary data.

Williamson et al. (1995), Heagerty and Zeger (1996), and Lumley (1996) had the appealing idea of, instead, basing the correlations on a natural association model for ordinal data for the two-way tables that cross-classify pairs of responses. For example, one could assume a uniform association model for global odds ratios for each pair of responses. In an exchangeable form, this requires only a single global odds ratio parameter to specify the working correlation structure. The association parameter can itself be modeled in terms of explanatory variables, to describe more precisely how the correlation may vary.

Williamson et al. (1995) illustrated their approach for the diabetic retinopathy example described in Section 9.1.3. The analysis in that section used ML methods. In their GEE analysis, the global odds ratio seemed to vary as a function of gender and the number of doses of insulin taken per day. They used the fit of the association model specifying a uniform global odds ratio at each fixed setting of gender and doses of insulin to generate the working correlation matrix and hence the GEE analyses. The GEE estimates and SE values they obtained were similar to those reported in Table 9.2 using ML.

9.2.5 Dealing with Missing Data

For any modeling with longitudinal data, missing data are often problematic. Inference using ML has the advantage of being applicable under weaker assumptions about the missing data mechanism than GEE requires. The GEE method requires the strong assumption that the data are *missing completely at random* (MCAR). This means that the missing data are a random sample of all observations. In particular, the probability that any observation is missing is independent of the value

of that observation. In modeling an ordinal arthritis assessment as a function of age, if the probability that the arthritis assessment is missing is the same for all subjects regardless of their age or their actual arthritis assessment, the data are MCAR.

ML methods are valid under the weaker assumption that the missing observations are *missing at random* (MAR). This means that given the observed data, the missingness mechanism does not depend on the unobserved data. That is, what caused the data to be missing does not depend on the data itself. If the probability that the arthritis assessment is missing varies according to the age of the subject but does not vary according to the arthritis assessment of subjects with the same age, the data are MAR.

Suppose that we divide the subjects into the group having a complete set of observations and the group having missing observations. If the data are MCAR, both groups should be random samples of the same population distribution of observations. If the data are MAR, both groups should be random samples of the same distribution of arthritis assessment within each level of age, but they do not have the same distribution of age [see Little and Rubin (2002) for further details about these definitions].

Likelihood-based inference is valid under the MAR condition. Under that condition, it is not necessary to specify the missing data mechanism. When the data are MAR but not MCAR, it is necessary with the GEE method to specify the missing data mechanism to deal with the potential bias. For ways of dealing with missing data, see Kenward et al. (1994), Mark and Gail (1994), Molenberghs et al. (1997), Little and Rubin (2002), Molenberghs and Verbeke (2005), and Kaciroti et al. (2006). Sometimes, information may be missing in a key covariate rather than the responses. Toledano and Gatsonis (1999) modified ordinal GEE equations to account for this.

9.3 TRANSITIONAL ORDINAL MODELING, GIVEN THE PAST

For a sample of patients with insomnia problems, Table 9.5 shows results of a randomized, double-blind clinical trial comparing an active hypnotic drug with a placebo. The response is the patient's reported time (in minutes) to fall asleep after going to bed. Patients responded before and following a two-week treatment period. The two treatments, active drug and placebo, form a binary explanatory variable. The subjects were randomly allocated to the treatment groups, so observations for the two treatment groups are independent samples. Here each subject forms a matched-pair type of cluster, with the two observations in a cluster being the ordinal response at the two occasions.

We could use marginal models to analyze the data. Table 9.6 displays sample marginal distributions for the four treatment × occasion combinations. At the initial occasion, the marginal distributions for the two treatments were similar, as was expected because of the random assignment of subjects to the treatment groups. From the initial to follow-up occasion, the time to falling asleep tended to shift

TABLE 9.5. Time to Falling Asleep, by Treatment and Occasion

Treatment	Initial	Time to Falling Asleep			
		Follow-up			
		<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

Source: S. F. Francom et al. (1989), with permission of John Wiley & Sons, Ltd.

TABLE 9.6. Sample Marginal Distributions of Table 9.5 for Response on Time to Falling Asleep

Treatment	Occasion	Response			
		<20	20–30	30–60	>60
Active	Initial	0.101	0.168	0.336	0.395
	Follow-up	0.336	0.412	0.160	0.092
Placebo	Initial	0.117	0.167	0.292	0.425
	Follow-up	0.258	0.242	0.292	0.208

downward for both treatments. The degree of shift seems greater for the active drug, indicating possible interaction, as is verified by model fitting (Agresti 2002, p. 469). Here we'll use an alternative type of model for these data, applying logits only to the follow-up response and treating the initial response as a covariate.

9.3.1 Comparisons that Control for Initial Response with Matched Pairs

Let Y_t denote the response at occasion t ($t = 1$, initial, $t = 2$, follow-up) and let x denote the treatment (0 = placebo, 1 = active drug). So Y_2 denotes the follow-up response, for treatment x with initial response Y_1. With scores assigned to the categories for the initial outcome, the model

$$\text{logit} [P(Y_2 \leq j)] = \alpha_j + \beta_1 x + \beta_2 y_1 \qquad (9.2)$$

controls for that initial response. In this model, the parameter β_1 compares the follow-up distributions on Y_2 for the treatments, controlling for initial observation y_1. This is an analog of an analysis of covariance model, with ordinal rather than

continuous response. This cumulative logit model refers to a univariate response (Y_2) rather than to the marginal distributions of a bivariate response (Y_1, Y_2).

In some situations, whether an effect of a certain type exists may differ between this type of model and a marginal model. For example, consider data in the form of Table 9.5, with responses at two times for two treatment groups. Suppose that the true marginal distributions for initial response are identical for the treatment groups, as we expect with random assignment of subjects to the groups. Suppose also that there is no treatment effect, in the sense that conditional on the initial response, the follow-up response distribution is identical for the treatment groups. Then the follow-up marginal distributions are also identical. By contrast, suppose that the initial marginal distributions are not identical, as might well happen with observational data for which randomization of subjects is not possible. Then, even when the conditional distributions for follow-up response are identical for the two treatment groups, the difference between follow-up and initial marginal distributions may differ between the treatment groups. In such cases it may be more informative to construct models that compare the follow-up responses while controlling for the initial response. Model (9.2), which does this, is an example of a *transitional model*.

9.3.2 Transitional Models with Time-Series Data

In longitudinal studies with relatively long time-series data, the focus is often on the dependence of observation Y_t on the responses observed previously $(y_1, y_2, \ldots, y_{t-1})$ as well as on the explanatory variables. Time-series models that include past observations as predictors are transitional models. A *Markov chain* is a transitional model for which, for all t, the conditional distribution of Y_t given $y_1, \ldots, y_{t-1}$ is assumed identical to the conditional distribution of Y_t given y_{t-1} alone. That is, given y_{t-1}, Y_t is conditionally independent of $Y_1, \ldots, Y_{t-2}$. Knowing the most recent observation, information about observations before that one does not help us predict the next observation. Many transitional models have Markov chain structure for at least part of the model. See, for example, Lindsey et al. (1997), Böckenholt (1999), and Müller and Czado (2005).

Transitional models can also include explanatory variables other than past observations. With k such explanatory variables, we could specify an ordinal model for each t, such as

$$\text{logit} \left[P(Y_t \leq j) \right] = \alpha_j + \beta y_{t-1} + \beta_1 x_{1t} + \cdots + \beta_k x_{kt}.$$

This model also permits an explanatory variable to take a different value for each t. For example, in a longitudinal medical study, a subject's values for predictors such as blood pressure and cholesterol level would be time varying. Kedem and Fokianos (2002, p. 99) used a cumulative logit transitional model of this form in which the explanatory variable is a periodic function of time and y_{t-1} is represented by indicators for the categories. Given the predictor values at each t, if we treat the observations by a subject as independent, this type of model can be fitted

with ordinary software. A higher-order Markov model could also include y_{t-2} and possibly other previous observations in the linear predictor.

In transitional models, the interpretation and magnitude of $\hat{\beta}$ depends on how many previous observations are in the model. Within-cluster effects often diminish markedly by conditioning on previous responses. This is an important difference from marginal models, for which the interpretation does not depend on the specification of the dependence structure. In some applications, it is more relevant to estimate the effects on Y_t without conditioning on previous response values. In addition, many transitional models that have been proposed have the limitations that subjects must be observed at the same times and an observation cannot be included in the fitting process if the previous observation(s) is missing.

9.3.3   Example: Insomnia Clinical Trial

The transitional type of model can be especially useful for matched-pairs data. Marginal models evaluate how the marginal distributions of Y_1 and Y_2 depend on explanatory variables. By contrast, a transitional model treats Y_2 as a univariate response, evaluating effects of explanatory variables while controlling for the initial response y_1.

Consider the insomnia study of Table 9.5. In model (9.2) we use scores (10, 25, 45, 75) for the four categories of the initial time to fall asleep y_1. This initial response y_1 plays the role of an explanatory variable, in addition to the treatment group predictor. Applying software for ordinary cumulative logit models to the univariate response Y_2, the ML treatment effect estimate is $\hat{\beta}_1 = 0.885$ (SE = 0.246). This provides strong evidence that follow-up time to fall asleep is lower for the active drug group. For any given value for the initial response, the estimated odds of falling asleep by a particular time for the active treatment are $\exp(0.885) = 2.4$ times those for the placebo group. In Exercise 9.6 we consider alternative analyses for these data.

CHAPTER NOTES

Section 9.1: Marginal Ordinal Modeling with Explanatory Variables

9.1. Much of the ML marginal modeling literature also provides models for the joint distribution, as in the retinopathy example of Williamson and Kim (1996) presented in Section 9.1.3. In an early application of ML for marginal modeling, Dale (1986) used cumulative logit models for the margins while using the global odds ratio to describe the joint distribution. Lang and Agresti (1994) gave other examples of simultaneous modeling of marginal and joint distributions. The ML approach was extended by Molenberghs and Lesaffre (1994), Heagerty and Zeger (1996), Molenberghs et al. (1997), and Lesaffre et al. (1998). For other uses of ML with marginal models for ordinal data, see Glonek and McCullagh (1995), Kim (1995), Glonek (1996), Lang and Eliason (1997), Lang et al. (1999), Bartolucci et al. (2001), Colombi and Forcina (2001), Vermunt et al. (2001), and

Ekholm et al. (2003). By contrast, Jokinen et al. (2006) focused on modeling the association structure for clustered ordinal data using various latent variable and Markov models. Rather than using odds ratios, they and Ekholm et al. (2003) described association using *dependence ratios*, which are measures of the form $P(Y_{is} = a, Y_{it} = b)/P(Y_{is} = a)P(Y_{it} = b)$.

Section 9.2: Marginal Ordinal Modeling: GEE Methods

9.2. For GEE methods for ordinal responses, see Heagerty and Zeger (1996), Miller et al. (1993), Lipsitz et al. (1994), Williamson et al. (1995), Lumley (1996), Ten Have et al. (1998), Miller et al. (2001), Huang et al. (2002), Parsons et al. (2006), Nores and Diaz (2008), Parsons et al. (2009), and references in Agresti and Natarajan (2001) and Liu and Agresti (2005). An earlier approach formed a weighted combination of estimates from separate models fitted to margins, using an empirically generated covariance matrix of the separate estimates (Stram et al. 1988). More general models with ordinal responses allow for dispersion parameters that also depend on covariates, as in Section 5.4. Toledano and Gatsonis (1996) used such models for estimating ROC curves from multiple interpretations of the same diagnostic study. Stiger et al. (1999) presented tests of the proportional odds assumption for GEE analyses of cumulative logit models of that form. Heagerty and Zeger (2000a) used multivariate continuation-ratio logit models. Williamson and Lee (1996) developed GEE methods for a mixture of an ordinal with a nominal response, modeling the marginal distributions and using odds ratios that are cumulative in the ordinal variable to describe the association.

9.3. Koch et al. (1977) used weighted least squares (WLS) to fit marginal models to categorical data. The WLS approach is simpler than ML but has severe limitations, such as needing categorical covariates and nonsparse marginal tables. Because of this, it is now rarely used in the form originally proposed, but it can be regarded as a natural predecessor of the GEE approach. In particular, Miller et al. (1993) showed that under certain conditions the solution of the first iteration in the GEE fitting process gives the WLS estimate. This equivalence uses initial estimates based directly on sample values and assumes a saturated association structure having a separate correlation parameter for each pair of response categories and each pair of observations in a cluster. In this sense, GEE (like ML) is an iterated form of WLS. Moreover, in this case, the covariance matrix for the estimates is the same with WLS and GEE approaches.

9.4. For designs with longitudinal observations of ordered categorical data, Brunner and Langer (2000) proposed a nonparametric model for marginal distributions to analyze treatment effects and interactions. The proposed methods also provide extensions of the Wilcoxon–Mann–Whitney test to factorial designs. See also the text by Brunner et al. (2002) and references in Section 7.7.2.

Section 9.3: Transitional Ordinal Modeling, Given the Past

9.5. Lindsey et al. (1997) proposed a simple Markov model, conditioning on the previous response using continuation-ratio logits. Böckenholt (1999) presented

mixed Markov cumulative probit models that take into account both within- and between-subject variability. Kosorok and Chao (1996) considered a Markov model for an ordinal response in continuous time. It is also possible to combine marginal models with transitional models. Lee and Daniels (2007) proposed such models for the analysis of longitudinal ordinal data. They introduced a proportional odds model of cumulative logit form for the marginal distributions and a Markov transition structure for the temporal dependence, using ML for model fitting. In related work using a variety of ordinal structures, Bartolucci and Farcomeni (2009) included a time-varying random effect that follows a Markov process. Albert (1994), Lindsey and Kaufmann (2004), and Kaciroti et al. (2006) proposed transitional models for ordinal time-series data. Varin and Vidoni (2006) considered a cumulative probit time-series model for which an underlying latent variable model has autoregressive structure. For a latent variable Y_t^* for which categorization into intervals provides the observed ordinal response, the model has the form

$$Y_t^* = \alpha + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \gamma_1 Y_{t-1}^* + \cdots + \gamma_p Y_{t-p}^* + \epsilon_t,$$

where $\{\epsilon_t\}$ are independent normal variates. Model fitting is complex, and Varin and Vidoni proposed a pseudolikelihood model-fitting method based on the *composite likelihood* approach that uses contributions to the likelihood function for pairs of observations.

EXERCISES

9.1. A study investigates home Internet use, measured as (none, <1 hour a day, >1 hour a day), for subjects in families living in a rural location and families living in an urban area. Each cluster consists of the people in a particular family. For this study, give an example of a (a) within-cluster effect and (b) between-cluster effect. Explain why you would expect the within-cluster correlation to be positive, and in each case, explain how the actual standard errors would compare to those obtained with the same number of independent observations.

9.2. In Section 8.2.2 we used ML to fit a marginal cumulative logit model to GSS responses on how well the government provides health care for the sick and protects the environment. Obtain GEE estimates, compare to the ML estimates, and interpret.

9.3. Analyze the crossover data in Table 8.5 using the GEE method with a marginal cumulative logit model. Compare $\{\hat{\beta}_t\}$ and their SE values to those obtained using ML in Section 8.4.3.

9.4. Refer to Exercise 8.5. Fit a marginal model to these data about government spending, using ML or GEE. Interpret the results.

9.5. Analyze the soft-drink comparison data of Table 8.8 in Section 8.6.3 using methods of this chapter that treat the three evaluations by a subject as potentially correlated. Interpret the results.

9.6. Refer to Table 9.5 on the clinical trial for insomnia patients.

    (a) To compare effects while controlling for the initial response, add an interaction term to model (9.2). Summarize how the estimated treatment effect varies according to the initial responses by showing that the estimated treatment log odds ratio changes from 0.00 to 1.41 as the initial response score goes from 10 to 75.

    (b) Now treat the initial response as qualitative, using indicator variables. Fit the model without interaction. Show that the estimated treatment log odds ratio is 0.911 (SE = 0.249), and interpret. Now fit the model with interaction terms. Explain why the results suggest that the active treatment seems relatively more successful at the two highest initial response levels.

    (c) Using ML or GEE, fit a marginal model for time to fall asleep, with predictors the treatment, the occasion, and treatment × occasion interaction. Interpret results.

CHAPTER 10

Clustered Ordinal Responses: Random Effects Models

In this chapter we present the alternative model type for clustered ordinal data, which contains a *random effect* term in the linear predictor for each cluster. Such models permit heterogeneity in response probabilities for the various clusters that have a particular setting of explanatory variables. In some studies, variability among random effects might represent heterogeneity caused by the absence from the model of certain explanatory variables that are associated with the response. Another use of a random effect term is to represent random measurement error in the explanatory variables.

A model with random effects is a type of latent variable model. The observed clusters are regarded as being sampled randomly from the set of all possible clusters, and the random effect for a cluster is an unobserved random variable. Conditional on the random effect, the observations in a cluster are treated as independent, whereas marginally, ignoring the random effects, they are associated. Early applications of such models for ordinal variables were in the context of extending factor analysis to categorical data, such as by Samejima (1969) and Bartholomew (1980, 1983).

In Sections 8.2.4 and 8.4.2 we introduced cluster effects in models for matched pairs and for matched sets. Such models have *conditional* interpretations for the effects of the explanatory variables, in the sense that those effects are conditional on the cluster. The effects are called *cluster specific*, or *subject specific* when each cluster is a subject. This contrasts with the marginal models presented in Chapter 9, which have *population-averaged* interpretations because effects are averaged over all the clusters.

In Section 10.1 we introduce ordinal response models with random effects and discuss model interpretation and inference. In Section 10.2 we present examples of models for which the random effect plays the role of an intercept term that varies among clusters. There we revisit examples analyzed with marginal models

in Chapters 8 and 9. In Section 10.3 we present examples of models with multiple random effect terms, which permit effects of explanatory variables as well as intercepts to vary among clusters. In Section 10.4 we discuss multilevel (hierarchical) models in which random effects enter at various levels, such as occurs in educational applications that include random effects for students as well as for schools. In the final section we discuss relevant issues in choosing between a random effects model, a marginal model, and some other type of model, such as a transitional model.

10.1   ORDINAL GENERALIZED LINEAR MIXED MODELS

The *generalized linear mixed model* (GLMM) is an extension of the generalized linear model that permits random effects as well as fixed effects in the linear predictor. In this chapter we introduce GLMMs for ordinal responses that are assumed to have a multinomial distribution at each particular value of the fixed and random effects. Such models are special cases of a multivariate generalized linear mixed model (Tutz and Hennevogl 1996).

10.1.1   Cumulative Logit Random Intercept Model

Let y_{it} denote the response for observation t in cluster i. Let $x_{1it}, x_{2it}, \ldots, x_{kit}$ denote the values of the k explanatory variables for that observation. Denote the random effect for cluster i by u_i. In the simplest and most common case, u_i is an intercept term in the model. The model is then called a *random intercept* model.

For outcome categories $j = 1, 2, \ldots, c - 1$, the cumulative logit model of proportional odds form with a random intercept is

$$\text{logit}[P(Y_{it} \le j)] = u_i + \alpha_j + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit}. \tag{10.1}$$

This model takes the linear predictor from the marginal model (9.1) and adds a random effect u_i to the intercept term α_j. It uses the same random effect for each cumulative probability. Using an overall intercept term of form $u_i + \alpha_j$ is also a way of allowing subjectivity in subjects' choices for outcome categories by allowing the ordinal scale cutpoints to vary among subjects (Farewell 1982; Wolfe and Firth 2002).

A subject with a relatively large positive (negative) u_i has relatively large (small) cumulative probabilities, and hence a relatively high (low) chance of occurring at the low end of the ordinal scale. We can express the model alternatively as

$$\text{logit}[P(Y_{it} \le j)] = \alpha_j - (u_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit}),$$

which naturally results from an underlying latent variable model. For this parameterization, increasing values of random and fixed effects correspond to increasing probabilities at the high end of the ordinal scale. Another way this interpretation

results is from replacing $P(Y_{it} \leq j)$ in (10.1) by $P(Y_{it} > j)$. The cutpoint parameters satisfy $\alpha_1 < \alpha_2 < \cdots < \alpha_{c-1}$, to reflect the ordering of cumulative probabilities for each i.

In practice, a random effect u_i is unobserved, so its value is unknown. It is usually assumed to vary from cluster to cluster according to a normal $N(0, \sigma_u^2)$ distribution. The *variance component* σ_u^2 is a parameter that is estimated together with the fixed effects. Why not treat the $\{u_i\}$ as fixed effects, that is, as parameters rather than random effects? One reason is that usually a study has a large number of clusters (e.g., one for each subject), so the model would then contain too many parameters. Treating $\{u_i\}$ as random effects, we have only a single additional parameter (σ_u) in the model, describing their dispersion. Another reason is that if $\{u_i\}$ were treated as fixed, inferences would extend only to those clusters and not to the population of clusters they represent.

10.1.2 Other Ordinal Logit Random Intercept Models

The same linear predictor structure holds with other link functions for which a common effect for each logit is plausible. For example, Hartzel et al. (2001a,b) used it with adjacent-categories logits (ACLs),

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = j + 1)} = u_i + \alpha_j + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit}.$$

An ACL model is more useful than a cumulative logit model when we want descriptions to contrast probabilities of response in pairs of categories rather than above versus below various points on the response scale. The ACL model is equivalent to the baseline-category logit (BCL) model,

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = c)} = (\alpha_j + \cdots + \alpha_{c-1}) + (c - j)(u_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit}).$$

The effects in the BCL model are the multiple $c - j$ of those in the ACL model. Random effects models can also use continuation-ratio logits or an alternative cumulative link such as the probit or complementary log-log.

When used with a common fixed effect $\boldsymbol{\beta}$ and a common random effect u_i for each logit, cumulative link models and adjacent-categories logit models both describe location effects and imply stochastically ordered response distributions at different settings of predictors. They typically provide similar substantive conclusions about the statistical and practical significance of effects.

10.1.3 Positive Correlation Induced by Random Effects Variability

As the variance σ_u^2 of the random effects increases, the correlation $\text{Corr}(y_{it}, y_{is})$ between two observations within the same cluster also tends to increase. This type of correlation is called an *intraclass correlation*.

Consider the cumulative link model with a random intercept. The underlying regression model refers to a continuous latent variable y^*. The latent outcome for observation t in cluster i is

$$y_{it}^* = \alpha + \boldsymbol{\beta}' \mathbf{x}_{it} + u_i + \epsilon_{it}.$$

Suppose that $\{u_i\}$ are independent $N(0, \sigma_u^2)$ variates and $\{\epsilon_{it}\}$ are independent errors that are also independent of $\{u_i\}$ and have variance σ^2. Then, conditional on the explanatory variables and for each pair (t, s) in a cluster,

$$\text{Corr } (y_{it}^*, y_{is}^*) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} \tag{10.2}$$

(Skrondal and Rabe-Hesketh 2004, p. 51). This equals the proportion of the total residual variance that is due to the variability σ_u^2 in the random effect. The correlation is positive and increases as σ_u^2 increases, for fixed σ^2.

For this latent variable construction, σ is nonidentifiable because the model still holds if the latent outcomes are linearly rescaled. The cumulative probit model results when ϵ_{it} has a normal distribution, and it is then common to let $\sigma = 1$, corresponding to the standard normal distribution. On this scale, the variability σ_u^2 of the random effect is such that $\text{Corr}(y_{it}^*, y_{is}^*) = \sigma_u^2/(\sigma_u^2 + 1)$. For example, σ_u values of $(0, 1, 2, 3)$ correspond to correlation values of $(0, 0.50, 0.80, 0.90)$. This correlation corresponds to a polychoric correlation for the ordered categorical scale.

The cumulative logit model results from categorizing an underlying standard logistic latent variable. Then $\text{Corr}(y_{it}^*, y_{is}^*) = \sigma_u^2/[\sigma_u^2 + (\pi^2/3)]$, since $\sigma = \pi/\sqrt{3}$ for the standard logistic distribution. For example, σ_u values of $(0, 1, 2, 3, 4)$ correspond to correlation values of $(0, 0.23, 0.55, 0.73, 0.83)$. We'll see in Section 10.5.1 that increasing σ_u also increases the difference between sizes of corresponding fixed effects in random effects models and in marginal models.

As suggested by formula (10.2) for the correlation in terms of variance components, a model with a random intercept implies a *nonnegative* correlation between clustered observations. In the boundary case $\sigma_u^2 = 0$ of no between-cluster heterogeneity, the correlation disappears and the clustered observations behave as if they are independent observations.

10.1.4 Parameter Interpretations for Random Effects Models

A fixed effects parameter in a model with random effects has a conditional intepretation. It refers to the consequence of changing the value of an explanatory variable, for which the fixed effect is the coefficient, for a given value of the random effect and the other fixed effects. Those fixed effects are of two types. First, consider an explanatory variable that varies in value among observations in a cluster. For example, in a crossover study comparing T drugs, for each subject the drug taken varies from observation to observation in that subject's cluster of T observations. For such an explanatory variable, its coefficient refers to the effect on the response

of a within-cluster (e.g., subject-specific) 1-unit increase of that predictor. The effect of that explanatory variable is a *within-cluster* effect.

Second, consider an explanatory variable with constant value among observations in a cluster. When each cluster is a person, an example is the person's gender or race. For such an explanatory variable, its coefficient refers to the effect on the response of a *between-cluster* 1-unit increase of that predictor. An example is a comparison of females and males using an indicator variable and its coefficient. However, this fixed effect applies only when the random effect as well as other explanatory variables in the model take the same values in both groups, such as a male and a female with the same values for their random effects.

It is in this sense that random effects models are *conditional* models, as both within- and between-cluster effects apply conditional on the random effect value. By contrast, effects in marginal models are averaged over all clusters; that is, they are *population averaged*. Those effects do not refer to a comparison at a fixed value of a random effect.

How can we interpret the variability in effects that this model implies for clusters having different random effect values? Consider observation y_{it} for cluster i at a particular setting for predictor x_k and observation y_{hs} for cluster h at setting $x_k + 1$. Their log odds ratio for a cumulative logit model of form (10.1) is

$$\text{logit}[P(Y_{hs} \le j \mid u_h)] - \text{logit}[P(Y_{it} \le j \mid u_i)] = \beta_k + (u_h - u_i).$$

When $\sigma_u = 0$, β_k is the usual form of cumulative log odds ratio for a model without random effects. When $\sigma_u > 0$, β_k is the cumulative log odds ratio for two observations in the same cluster ($h = i$) or with the same random effect value. Otherwise, we cannot observe $u_h - u_i$, but the difference $u_h - u_i$ is a random variable having a $N(0, 2\sigma_u^2)$ distribution. Thus, $100(1 - \alpha)\%$ of these log odds ratios fall within

$$\beta_k \pm z_{\alpha/2}\sqrt{2}\sigma_u. \tag{10.3}$$

10.1.5 ML Model Fitting and Inference

The model-fitting process estimates the fixed effects and the standard deviation σ_u of the random effects that describes the variability among clusters. Hedeker and Gibbons (1994; 2006, Sec. 10.2.4), Best et al. (1996), Tutz and Hennevogl (1996), and Hartzel et al. (2001b) discussed model fitting for ordinal random effects models.

An ordinal GLMM can be regarded as a two-stage model. At the first stage, conditional on the random and fixed effects, observations are assumed to be independent, as in an ordinary multinomial model. At the second stage, the random effects are assumed to be independent realizations from a normal distribution. Integrating out the random effects gives a marginal distribution for the response outcomes and a *marginal likelihood function*. This is a function of the fixed effects parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\alpha_1, \ldots, \alpha_{c-1}, \beta_1, \ldots, \beta_k\}$ and the parameter σ_u of the $N(0, \sigma_u^2)$ random effects distribution. Averaged with respect to the distribution

of the $\{u_i\}$, the model implies nonnegative correlation among observations within a cluster, as discussed in Section 10.1.3.

Here are some details for a cumulative link model, adding a random intercept to the basic model (5.1) with link function h: For each of the T_i observations in cluster i, let the response y_{it} for observation t be represented by a vector $\mathbf{y}_{it}$ of c binary indicators. That is, let $y_{ijt} = 1$ if the observation falls in category j and let $y_{ijt} = 0$ otherwise, $i = 1, \ldots, n$, $j = 1, \ldots, c$, $t = 1, \ldots, T_i$. Given u_i, let $\pi_j(\mathbf{x}_{it}; u_i) = P(y_{ijt} = 1)$. We treat $\mathbf{y}_{it}$ as a multinomial observation with probability mass function

$$\pi_1(\mathbf{x}_{it}; u_i)^{y_{i1t}} \pi_2(\mathbf{x}_{it}; u_i)^{y_{i2t}} \cdots \pi_c(\mathbf{x}_{it}; u_i)^{y_{ict}}.$$

Each term in this product is a difference of cumulative probabilities with the inverse link function,

$$\pi_j(\mathbf{x}_{it}; u_i) = h^{-1}(u_i + \alpha_j + \boldsymbol{\beta}'\mathbf{x}_{it}) - h^{-1}(u_i + \alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_{it}),$$

with $\alpha_0 = -\infty$ and $\alpha_c = \infty$. With a $N(0, \sigma_u^2)$ probability density function for u_i, the marginal likelihood function has the form

$$\prod_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \prod_{j=1}^{c} \left(h^{-1}(u_i + \alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta}) - h^{-1}(u_i + \alpha_{j-1} + \mathbf{x}_{it}'\boldsymbol{\beta}) \right)^{y_{ijt}} \right] \left(\frac{1}{\sqrt{2\pi}\sigma_u} e^{-u_i^2/2\sigma_u^2} \right) du_i \right\}.$$

The main computational difficulty in fitting GLMMs is the need to evaluate this integral to obtain the marginal likelihood function. The integral does not have a closed form. Numerical methods for approximating the marginal likelihood function can be computationally intensive, especially for models with multiple random effect terms. Once the marginal likelihood function is approximated, standard methods such as the Newton–Raphson algorithm can maximize it, yielding ML estimates of parameters. As a by-product, inverting the approximated observed information matrix provides a large-sample covariance matrix for ML estimates. Inference about fixed effects then proceeds in the usual way. For example, likelihood-ratio tests can compare nested models. Asymptotics for the model apply as the number of clusters increases rather than as the numbers of observations within the clusters increase.

For relatively simple GLMMs such as random intercept models, *Gauss–Hermite quadrature* approximates the integral that determines the marginal likelihood function by a finite weighted sum that is evaluated at certain *quadrature points*. Essentially, the normal density is approximated by a discrete histogram with bars centered at the quadrature points. The approximation improves as the number q of quadrature points increases. Similarly, as q increases, subsequent approximations for the ML parameter estimates and their SE values improve. An adaptive version of

Gauss–Hermite quadrature centers the quadrature points with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. This improves the efficiency, reducing the number of quadrature points needed for an effective approximation. The number of unique clusters is the biggest factor in determining the amount of time that the fitting process requires. Model fitting can be rather slow for large data sets with continuous covariates.

Gauss–Hermite quadrature can handle most models used in practice. With a complex random effect structure, however, it may not be feasible. *Monte Carlo* methods simulate in order to approximate the integral that determines the marginal likelihood function. The Gauss–Hermite and Monte Carlo methods approximate the ML parameter estimates but converge to the ML estimates as they are applied more finely: for example, as the number of quadrature points increases for numerical integration. This is preferable to other approximate methods that are simpler but need not yield estimates near the ML estimates. Such approaches, such as *Laplace approximation* and *penalized quasi likelihood* (PQL), replace the function to be integrated by an approximation for which the integral has closed form (e.g., Keen and Engel 1997; Hartzel et al. 2001b). These methods can perform poorly relative to ML, especially when the true variability σ_u of the random effects is large. For example, the PQL method can substantially underestimate σ_u.

Another approach to model fitting is Bayesian. With it, the distinction between fixed and random effects no longer occurs. A prior distribution is assumed for each effect of either type. We discuss this approach in Chapter 11.

10.1.6 Prediction and Inference About Random Effects

The model-fitting process can also provide predictions $\{\hat{u}_i\}$ for the random effects. The prediction $\hat{u}_i$ is the expected value of its posterior conditional distribution, given the data. Calculation of a predicted value itself requires numerical integration or Monte Carlo approximation. Similarly, it is possible to predict the value of a cumulative probability for a particular cluster. One averages the expression for the estimated cumulative probability, as a function of the ML parameter estimates and random effect, with respect to the posterior conditional distribution of the random effect. A simpler approach substitutes $\hat{u}_i$ into the estimated cumulative probability. This gives a somewhat different value, though, because the nonlinearity implies that the expected value of the cumulative probability is not the same as the cumulative probability evaluated at the expected random effect value.

The predictions $\{\hat{u}_i\}$ induce corresponding predictions for effects of interest, such as cumulative odds ratios. These predictions exhibit shrinkage relative to those that use only the sample data in the specific cluster. Shrinkage estimators can be far superior to sample values when the sample size for estimating each parameter is small, when there are many parameters to estimate, and when the true parameter values are roughly equal.

Sometimes it is also relevant to conduct inference about the standard deviation σ_u of $\{u_i\}$, which describes the variability among clusters. One such inference compares the model with its special case in which $\sigma_u = 0$. The simpler model

then falls on the boundary of the parameter space relative to the more complex model, since σ_u cannot be negative. When this happens, the usual likelihood-ratio chi-squared test for comparing models is not valid. Likewise, a Wald statistic such as $\hat{\sigma}_u/\text{SE}$ does not have an approximate standard normal null distribution. (When $\sigma_u = 0$, because $\hat{\sigma}_u < 0$ is impossible for an ML estimate, $\hat{\sigma}_u$ is not approximately normally distributed around σ_u.) For testing $H_0: \sigma_u = 0$ against $H_a: \sigma_u > 0$ for a random intercept model, the asymptotic null distribution of the likelihood-ratio statistic has probability $\frac{1}{2}$ at 0 and $\frac{1}{2}$ following the shape of a chi-squared distribution with df $= 1$. The test statistic value of 0 occurs when $\hat{\sigma}_u = 0$, in which case the maximum of the likelihood function is identical under H_0 and H_a. When $\hat{\sigma}_u > 0$ and the observed test statistic equals t, the P-value for this large-sample test is $\frac{1}{2}P(\chi_1^2 > t)$, half the P-value that applies for χ_1^2 asymptotic tests.

10.2   EXAMPLES OF ORDINAL RANDOM INTERCEPT MODELS

To illustrate ordinal models with random effects and their interpretation, we show three examples. The first is the crossover study analyzed in Section 8.4.3, which has a within-cluster fixed effect (the treatments) and a random effect (the subject term). The second example, the arthritis study analyzed in Section 9.2.3, also has between-subject predictors. The third example shows that ordinal models are also useful for analyzing count responses that take relatively few values and do not satisfy a standard distribution for counts such as the Poisson. We finish by describing a generalization of the random intercept model for ordinal time-series data that allows observations closer together in time to be more highly correlated.

10.2.1   Example: Crossover Study for Treating Dysmenorrhea

Table 8.5 in Section 8.4.3 showed the results of a three-period crossover study designed to compare placebo (treatment A) with a low-dose analgesic (treatment B) and a high-dose analgesic (treatment C) for relief of dysmenorrhea, using response categories (no relief, moderate relief, complete relief). Each cluster is a particular woman's set of observations on the three treatments. In Section 8.4.3 we analyzed the data with a marginal cumulative logit model that is designed to detect location differences among the treatments.

Let y_{it} denote the outcome on treatment t for subject i. Let x_{it} be a dummy variable that equals 1 when subject i uses treatment t and equals 0 otherwise. In this context, cumulative logit model (10.1) is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_t x_{it}, \qquad j = 1, 2, \quad t = 1, 2, 3.$$

The random intercept u_i is assumed to vary randomly among subjects according to a $N(0, \sigma_u^2)$ distribution. For identifiability we set $\beta_1 = 0$ for treatment A, the

TABLE 10.1. Results for Models Fitted to Table 8.5, with SE Values in Parentheses

Comparison	Cumulative Logit Marginal	Cumulative Logit Random Effects	Adjacent-Cat. Logit Random Effects
$\hat{\beta}_B - \hat{\beta}_A$	2.038 (0.360)	2.030 (0.326)	1.460 (0.248)
$\hat{\beta}_C - \hat{\beta}_A$	2.430 (0.372)	2.410 (0.331)	1.724 (0.255)
$\hat{\beta}_C - \hat{\beta}_B$	0.392 (0.252)	0.380 (0.284)	0.265 (0.196)

placebo. So the model states that

$$\text{logit}[P(Y_{i1} \le j)] = u_i + \alpha_j, \qquad \text{logit}[P(Y_{i2} \le j)] = u_i + \alpha_j + \beta_2,$$

$$\text{logit}[P(Y_{i3} \le j)] = u_i + \alpha_j + \beta_3, \qquad j = 1, 2.$$

For $H_0 : \beta_2 = \beta_3 = 0$, the likelihood-ratio test statistic is double the difference between the maximized log-likelihood for this model and for the simpler null model having $\beta_2 = \beta_3 = 0$. This equals 68.4, with df = 2. Table 10.1 shows the ML estimated effects, but replacing β_t in this model by $-\beta_t$ so that larger values of β_t correspond to more positive results (e.g., a greater probability of complete relief), as in the parameterization for the marginal cumulative logit model in Section 8.4.3. The contrasts of $\{\hat{\beta}_t\}$ provide the same conclusions as does marginal model analysis. Treatments B and C clearly differ from placebo, but there is only weak evidence that the high dose is better than the low dose. The fixed effects estimates have subject-specific log odds ratio interpretations. For a given subject, for instance, the estimated odds that relief for the low-dose analgesic is moderate or complete rather than none, or complete rather than moderate or none, are $\exp(2.41) = 11.1$ times the estimated odds for the placebo.

The random effects have $\hat{\sigma}_u = 0.0$. So $\{\hat{\beta}_t\}$ and their SE values are the same as if the three responses for a woman were treated as independent and we fitted an ordinary cumulative logit model of proportional odds form. When $\sigma_u = 0$, the cumulative logit random effects model also implies the cumulative logit marginal model with the same parameter values and with independent observations. Table 10.1 also shows ML results for the cumulative logit marginal model fitted in Section 8.4.3. The ML estimates and SE values for the marginal model are not exactly the same as for the random effects model, because we fitted the marginal model in Section 8.4.3 while allowing the three observations to be correlated. (The random effects model with $\hat{\sigma}_u = 0$ corresponds to the marginal model fit treating the three responses as independent.) However, results are quite similar for the two models.

The corresponding adjacent-categories logit random effects model is

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = j + 1)} = u_i + \alpha_j + \beta_t x_{it}, \qquad j = 1, 2, \quad t = 1, 2, 3.$$

This model is equivalent to the baseline-category logit model,

$$\log \frac{P(Y_{it} = 1)}{P(Y_{it} = 2)} = u_i + \alpha_1 + \beta_t x_{it}, \quad \log \frac{P(Y_{it} = 1)}{P(Y_{it} = 3)} = 2u_i + (\alpha_1 + \alpha_2) + 2\beta_t x_{it}$$

for $t = 1, 2, 3$. Table 10.1 also shows ML results for this model. As expected, estimates are somewhat smaller than for the cumulative logit model, because the $\{\beta_t\}$ refer to local log odds ratios rather than cumulative log odds ratios. However, comparing their size relative to their SE values yields substantive results similar to those for the cumulative logit model.

10.2.2 Example: Arthritis Clinical Trial

Section 9.2.3 used a cumulative logit marginal model to analyze data from a randomized clinical trial comparing a drug (auranofin) with a placebo for the treatment of rheumatoid arthritis. At the start of the study and after one month, three months, and five months of treatment, patients assessed their arthritis on the ordinal scale (good, fair, poor). Other explanatory variables measured were age and sex.

We now analyze the data with a cumulative logit random effects model. As in the marginal model analysis, we treat the baseline self-assessment as a covariate. Let Y_{it} denote follow-up response t, $t = 1, 2, 3$, for subject i. Each subject forms a cluster for this model, with three observations in each cluster except for a few missing observations. The model is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_1 t_2 + \beta_2 t_3 + \beta_3 b + \beta_4 a + \beta_5 s + \beta_6 tr,$$

for covariates $b = $ baseline response and $a = $ age and for dummy variables t_2 for time 2 and t_3 for time 3, s for sex ($1 = $ male, $0 = $ female), and tr for treatment ($1 = $ auranofin, $0 = $ placebo). ML fitting yields the prediction equation

$$\text{logit}[\hat{P}(Y_{it} \leq j)] = \hat{u}_i + \hat{\alpha}_j - 0.15 t_2 + 0.34 t_3 - 1.59 b - 0.01 a + 0.17 s + 0.84 tr$$

with $\hat{\sigma}_u = 1.92$.

To illustrate interpretation of a within-subjects effect, consider the coefficient 0.34 of t_3: For a given subject, the estimated odds of response "good" instead of "fair" or "poor" and the estimated odds of response "good" or "fair" instead of "poor" at time 3 were $\exp(0.34) = 1.40$ times the estimated time at time 1. To illustrate interpretation of a between-subjects effect, consider the coefficient 0.84 of tr: For a subject taking auranofin and a subject taking a placebo having the same random effects value, the estimated odds of response "good" instead of "fair" or "poor" and the estimated odds of response "good" or "fair" instead of "poor" for the auranofin patient were $\exp(0.84) = 2.3$ times the estimated odds for the placebo patient. Considering all random effect values for the two groups, since $z_{0.25} = 0.674$, 50% of the odds ratio values are estimated to fall within

$$\exp(\hat{\beta}_6 \pm 0.674\sqrt{2}\hat{\sigma}_u) = \exp[0.84 \pm 0.674\sqrt{2}(1.92)], \quad \text{which is } (0.37, 14.4).$$

The treatment effect shows considerable variability.

TABLE 10.2. Results for Models Fitted to Data from Arthritis Clinical Trial, with SE Values in Parentheses

Effect	Cumulative Logit Marginal	Cumulative Logit Random Effects	Adjacent-Cat. Logit Random Effects
Time 3	0.289 (0.098)	0.409 (0.161)	0.367 (0.139)
Baseline	−1.042 (0.136)	−1.600 (0.205)	−1.374 (0.185)
Treatment	0.554 (0.176)	0.850 (0.274)	0.753 (0.240)

When we constrain $\sigma_u = 0$, the maximized log-likelihood decreases by 72.1. This gives very strong evidence that $\sigma_u > 0$. The likelihood-ratio test statistic of $2(72.1) = 144.1$ has P-value $= \frac{1}{2}P(\chi_1^2 > 144.1)$, which is negligible. As in the marginal model case, effects are small and not statistically significant for time 2, age, and sex. A simpler model smooths the estimates of the substantively important effects by deleting the terms in the model for time 2, age, and sex. Table 10.2 shows the ML estimates from fitting this random effects model (again, it has $\hat{\sigma}_u = 1.92$). The table also shows the GEE estimates from fitting the corresponding marginal model, with independence working correlation structure. Results are substantively similar in the two cases. At the follow-up observations, the treatment group tends to have a better response.

From Table 10.2, ML estimates and standard errors are roughly 50% larger for the random effects model than for the marginal model. This reflects the fact that $\hat{\sigma}_u = 1.92$ for the random effects model is not near 0, reflecting substantial within-subject positive correlations among the repeated responses. The estimated intraclass correlation for an underlying latent variable equals $\hat{\sigma}_u^2/[\hat{\sigma}_u^2 + (\pi^2/3)] = 0.53$. In Section 10.5.1 we explain why cluster-specific effects tend to be larger than population-averaged effects, more so as $\hat{\sigma}_u$ increases.

Table 10.2 also shows ML results for the random effects model using adjacent-categories logits, for which $\hat{\sigma}_u = 1.66$. Estimated effects are a bit smaller than with the cumulative logit model, as expected, but results are substantively similar.

10.2.3 Example: Repeated Measures of Zero-Inflated Count Data

The next example differs from others in this book in that the response variable is an integer count rather than ordinal categorical. For count responses, sometimes the frequency of zero counts is much higher than expected with standard discrete models. For example, suppose that a longitudinal study observes each year the number of times that each subject made a medical appointment because of illness. Some subjects may have a zero observation in a particular year because of chance, whereas others may have a zero observation because of a doctor avoidance phobia or (in some countries) because of the cost and/or their lack of medical insurance. The data are then said to be *zero-inflated*.

Methods for modeling clustered zero-inflated count data include random effect models of various types, such as (a) a *hurdle model*, which uses logistic regression

to model whether an observation is zero or positive and a separate loglinear model with a truncated distribution for the positive counts; (b) a *zero-inflated Poisson model*, which for each observation uses a mixture of a Poisson loglinear model and a degenerate distribution at 0; and (c) a *zero-inflated negative binomial model*, which allows overdispersion relative to the zero-inflated Poisson model. Model (a) requires separate parameters for the effects of explanatory variables in the logistic model and in the loglinear model. Models (b) and (c) require separate parameters for the effects of explanatory variables on the mixture probability and in the loglinear model. It addition, those models can encounter fitting difficulties if there is zero deflation at any settings of the explanatory variables.

When the response variable has relatively few distinct count outcomes, a simple alternative approach applies a cumulative link random effects model to the count data (Saei et al. 1996). The first category is the zero outcome, and each other count outcome is a separate category. When the count can take a large number of values, the count oucomes are grouped into a set of ordered categories. It's then best to use at least four categories to avoid a substantial efficiency loss. This ordinal categorical approach has the advantage of a single set of parameters for describing effects. Those parameters describe effects overall rather than conditional on a response being positive.

Min and Agresti (2005) illustrated this approach with data from a pharmaceutical study comparing two treatments (labeled A and B) for a particular disease in terms of the number of episodes of a certain side effect observed at six times. The study had 118 patients, with half randomly allocated to each treatment.[1] Table 10.3 shows the frequencies of the side effect for the treatments, which took values between 0 and 6. About 83% of the observations were zeros, and Min and Agresti showed that there is strong evidence of zero inflation for standard models for counts such as a Poisson GLMM with a random intercept. The other explanatory variable was the time that had elasped since the preceding observation.

We group the response variable into five categories: 0, 1, 2, 3, 4, and > 4. For this grouped response, a cumulative logit model for observation t on subject i is

$$\text{logit}[P(Y_{it} \le j)] = u_i + \alpha_j + \beta_1 tr + \beta_2 \log(\text{time}), \qquad j = 0, \ldots, 4,$$

TABLE 10.3. Side-Effect Frequencies for Treatments A and B

Treatment	Frequency						
	0	1	2	3	4	5	6
A	312	30	11	0	1	0	0
B	278	39	20	6	7	2	2
Total	590	69	31	6	8	2	2

Source: Data supplied by Yongyi Min.

[1]The complete data set is at www.stat.ufl.edu/~aa/ordinal/ord.html.

where *tr* is an indicator for whether the subject uses treatment A. The model fit suggests that treatment A has a lower expected number of episodes than treatment B, as $\hat{\beta}_1 = 0.977$ has SE = 0.431. The estimated odds that the number of side effects falls below any fixed level with treatment A are $\exp(0.977) = 2.7$ times the estimated odds for treatment B. Time between visits has a negative effect on the cumulative logit and hence a positive effect on the expected number of episodes of the side effect, as expected, although it is not significant ($\hat{\beta}_2 = -0.153$, SE = 0.181). The estimate $\hat{\sigma}_u = 1.73$ (with SE = 0.25) of the variability among $\{u_i\}$ suggests considerable within-subject positive correlation among the repeated responses.

10.2.4 Autoregressive Structure for an Ordinal Time Series

The model fitting presented in Section 10.1.5 for a cumulative link model with a random intercept corresponds to a latent variable model

$$y_{it}^* = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + u_i + \epsilon_{it},$$

in which $\{u_i\}$ are independent $N(0, \sigma_u^2)$ variates and $\{\epsilon_{it}\}$ are independent variates (over i and t) with distribution having a standardized cdf whose inverse provides the link function. In practice, even when such a linear predictor is sensible, sometimes it is not realistic to assume that ϵ_{it} is uncorrelated with $\epsilon_{it'}$. Correspondingly, for the observed response y_{it}, it may not be realistic to assume that y_{it} and $y_{it'}$ are uncorrelated, conditional on u_i. An example is a longitudinal study for which each cluster consists of a long time series of observations. We then often expect observations within a cluster to be more strongly correlated when $|t - t'|$ is small than when $|t - t'|$ is large.

Varin and Czado (2010) provided a cumulative probit model for which $\{\epsilon_{it}\}$ have, instead, a Markov autoregressive structure. For this model, given $\epsilon_{i,t-1}$, ϵ_{it} is independent of $\epsilon_{i,t-2}, \epsilon_{i,t-3}, \ldots$. This structure implies that $\{\epsilon_{it}\}$ are correlated but with weaker correlations farther apart in time. A severe complication then is that the integral that determines the likelihood function is much more complex, since the joint distribution of the observations, given u_i, no longer factors into univariate components. Since ML is intractable with this model for large clusters, they proposed a pseudolikelihood model-fitting method based on the *composite likelihood* approach that uses contributions to the likelihood function for pairs of observations that are close together in time. As a consequence, the integrals that need to be evaluated are bivariate. This approach is much simpler for cumulative probit models than for other cumulative links, because (as we explain further in Section 10.5.1) averaging out the random effects yields a multivariate normal density function, so the pseudolikelihood approach then integrates bivariate normal densities. Varin and Czado illustrated their method with a longitudinal study on the determinants of migraine headache severity that recorded four observations per day on each subject for anywhere between 4 and 338 days.

10.3 MODELS WITH MULTIPLE RANDOM EFFECTS

Ordinal logit models with random intercepts such as the cumulative logit model (10.1) generalize to incorporate multiple random effects. The more general cumulative logit model of proportional odds form is

$$\text{logit}[P(Y_{it} \leq j)] = \alpha_j + \mathbf{u}_i' \mathbf{w}_{it} + \boldsymbol{\beta}' \mathbf{x}_{it}, \qquad j = 1, \ldots, c - 1. \tag{10.4}$$

The multivariate random effects $\mathbf{u}_i$ have their own explanatory variables. The random intercept models in Sections 10.1 and 10.2 had univariate $\mathbf{w}_{it} = 1$ and $\mathbf{u}_i = u_i$. The random effect $\mathbf{u}_i$ in (10.4) is usually assumed to have a multivariate normal distribution with unknown variances and correlations. In this section we present examples in which bivariate random effects are natural in ordinal logit models. A separate strand of research, not considered here, constructs models for a multivariate normal latent variable. See, for example, Todem et al. (2007).

10.3.1 Random Intercepts and Random Slopes

One important model of the form (10.4) has a random slope as well as a random intercept. Such a model allows an explanatory variable effect as well as the intercept to vary among clusters. For example, consider a randomized clinical trial that observes subjects initially on an ordinal response variable and then randomly assigns them to a treatment or control group and observes them at several follow-up occasions. Let y_{it} be the ordinal response for subject i at time t, and let x_i be an indicator for subject i that takes value 1 for treatment and 0 for control. Consider the model

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_1 x_i + (\beta_2 + v_i)t + \beta_3(x_i \times t),$$

where (u_i, v_i) has a bivariate normal distribution with means 0 and unknown standard deviations σ_u and σ_v and correlation ρ.

In this model, β_1 is the treatment effect initially (i.e., at $t = 0$), expected to be 0 because of randomization to the treatment and control groups. At follow-up times, the cumulative logit has linear trend over time with slope $\beta_2 + v_i$ for a subject from the placebo group and slope $\beta_2 + \beta_3 + v_i$ for a subject from the treatment group. That is, the model allows the trend over time to vary from subject to subject, with a mean of β_2 for the placebo group and a mean of $\beta_2 + \beta_3$ for the treatment group. Section 10.3.3 illustrates this type of model in the context of a psychiatric study.

10.3.2 Comparing Models with Differing Numbers of Random Effect Terms

A model with a random intercept u_i and a random slope v_i has three parameters associated with the random effects: σ_u and σ_v and $\rho = \text{Corr}(u_i, v_i)$. Suppose that you want to test whether the random component v_i of the slope effect can be dropped from the model, giving a model having the same slope in every cluster.

The simpler model has two fewer parameters, lacking σ_v and ρ. The asymptotic distribution of the likelihood-ratio statistic comparing the two models is a mixture of two chi-squared distributions, one with df $= 1$ and one with df $= 2$ (Molenberghs and Verbeke 2007).

More generally, to compare nested models with k versus $k + 1$ correlated random effects, the null distribution of the likelihood-ratio statistic is an equal mixture of chi-squared distributions with df $= k$ and df $= k + 1$. The P-value is the average of the right-tail probabilities above the observed test statistic value for the two distributions. The result mentioned in Section 10.1.6 for testing H_0: $\sigma_u = 0$ in a random intercept model is the special case with $k = 0$; in that case, the chi-squared distribution with df $= 0$ is degenerate at 0.

10.3.3 Example: Evaluating a Drug for Schizophrenia

In a psychiatry study, Hedeker and Gibbons (1994, 2006, Sec. 10.3) applied ordinal models having both a random intercept and a random slope. Patients suffering from schizophrenia were randomly assigned to receive a antipsychotic drug or placebo. The patients were then observed weekly for up to six weeks on an ordinal severity of mental illness scale: normal or borderline, mild or moderate, marked, severe. The data were also analyzed[2] by Rabe-Hesketh and Skrondal (2008, Chap. 7). Here we summarize some models and analyses, and the books cited provide further details.

Let x_i be an indicator variable for group ($1 = $ drug, $0 = $ placebo). For each group, the cumulative logits have approximately a linear trend over time when time is measured by the square root of the week number t for the observation. Over the six weeks, the proportion in the first category increased from 0.0 to 0.1 for the placebo group and from 0.0 to 0.35 for the drug group.

Suppose that we naively ignored the repeated measurement aspect of the study and treated the several observations for each patient as independent, using the model

$$\text{logit}[P(Y_{it} \leq j)] = \alpha_j + \beta_1 x_i + \beta_2 \sqrt{t} + \beta_3 (x_i \times \sqrt{t})$$

assuming independent multinomial observations. Table 10.4 shows some results for this marginal model. On the root-time scale, the time effect has estimated trend of $\hat{\beta}_2 = 0.54$ for the placebo group and $\hat{\beta}_2 + \hat{\beta}_3 = 1.29$ for the drug group. This model ignoring the clustering is inappropriate for obtaining SE values, and Table 10.4 does not report them.

To reflect the clustered nature of the data, we add a random intercept for each patient,

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_1 x_i + \beta_2 \sqrt{t} + \beta_3 (x_i \times \sqrt{t}).$$

The previous model is the special case in which $\sigma_u = 0$. The trend effects are now subject specific rather than population averaged. The time effect has estimated

---

[2]The data are at the web site www.stata-press.com/data/mlmus2.html.

TABLE 10.4. Results for Cumulative Logit Models Fitted to Data from Study Comparing Drug to Placebo for Treating Schizophrenia, with SE Values in Parentheses

Effect	Cumulative Logit Model		
	Marginal	Random Intercept	Random Intercept/Slope
Treatment ($\hat{\beta}_1$)	0.00	0.05 (0.31)	−0.11 (0.40)
Time ($\hat{\beta}_2$)	0.54	0.77 (0.12)	0.88 (0.24)
Interaction ($\hat{\beta}_3$)	0.75	1.21 (0.13)	1.72 (0.27)
$\hat{\sigma}$ for (u_i, v_i)	(0.0, 0.0)	(1.94, 0.0)	(2.67, 1.43)
Log-likelihood	−1878.1	−1701.4	−1662.8

Source: Results from Hedeker and Gibbons (2006) and Rabe-Hesketh and Skrondal (2008).

trend $\hat{\beta}_2 = 0.77$ for the placebo group and $\hat{\beta}_2 + \hat{\beta}_3 = 1.97$ for the drug group. Results are substantively similar but effects are larger because they are subject specific. (In Section 10.5.1 we explain why cluster-specific effects tend to be larger than population-averaged effects, more so as $\hat{\sigma}_u$ increases.) The estimate $\hat{\sigma}_u = 1.94$ reflects considerable heterogeneity among the patients in their propensity toward being mentally well.

To permit the slope of the time effect also to vary among patients, we use the model with both a random intercept and a random slope,

$$\text{logit}[P(Y_{it} \le j)] = u_i + \alpha_j + \beta_1 x_i + (\beta_2 + v_i)\sqrt{t} + \beta_3(x_i \times \sqrt{t}),$$

where (u_i, v_i) have a bivariate normal distribution with standard deviations (σ_u, σ_v). The foregoing model is the special case in which $\sigma_v = 0$. Table 10.4 shows that the time trend effect varies around an estimated mean of $\hat{\beta}_2 = 0.88$ for the placebo group and an estimated mean of $\hat{\beta}_2 + \hat{\beta}_3 = 2.60$ for the drug group, with standard deviations $\hat{\sigma}_v = 1.43$. The large ratio ($\hat{\beta}_3/\text{SE}$) provides strong evidence that the mean trend is greater for the drug than for placebo.

To analyze whether adding the random effects improved the fit, we compare maximized log-likelihood values for the three models. Rabe-Hesketh and Skrondal (2008, p. 301) reported that double the difference of log-likelihoods for the model with the random intercept and the simpler model having $\sigma_u = 0$ is 353. The asymptotic null distribution under H_0: $\sigma_u = 0$ is an equal mixture of χ_1^2 and degenerate at 0, so there is extremely strong evidence that the random intercept model is preferred. Double the difference of log-likelihoods for the model with the random intercept and random slope and the simpler model having $\sigma_v = 0$ is 77. The asymptotic null distribution under H_0: $\sigma_v = 0$, in which case $\text{Corr}(u_i, v_i) = 0$ also, is an equal mixture of χ_1^2 and χ_2^2. So there is extremely strong evidence that it is helpful to permit both random intercepts and random slopes. The estimated $\text{corr}(u_i, v_i)$ is −0.41. Those having less of a propensity to be mentally well tended to have a greater rate of improvement in their responses.

Table 10.4 shows that the SE values of the trend effects increase when we permit the slopes to vary among patients. This merely reflects that an estimate of a mean

effect has smaller SE when the effect is assumed to have no variability than when it is permitted to vary. The smaller SE for the simpler model can be misleading if the model fails to capture an important source of variability.

More generally, a model could allow (u_i, v_i) to have different covariance structure for the treatment and placebo groups. This would be useful, for example, if subjects in the placebo group have slopes that vary relatively little (perhaps about a small value or even 0) whereas subjects in the treatment group have slopes that vary substantially, with some subjects having essentially no change over time but others showing a rapid improvement. Additional useful analyses include predicting subject-specific cumulative probabilities or averaging over the estimated random effects distribution to estimate population-averaged effects for the implied marginal model. For details, see Hedeker and Gibbons (2006) and Rabe-Hesketh and Skrondal (2008).

10.3.4   Example: Heterogeneity in Multicenter Clinical Trials

Many applications deal with a comparison of two groups on an ordinal response for data stratified on a third variable. The data form several $2 \times c$ contingency tables. The goals include summarizing the association in the $2 \times c$ tables and analyzing whether and how that association varies among the strata.

The strata are sometimes themselves a sample, such as a sample of schools in a state or a sample of medical centers in a nation. A random effects approach is then natural, treating each stratum as a cluster. With a random sampling of strata, inferences can extend to the population of strata. The fit of the random effects model provides a simple summary such as an estimated mean and standard deviation of cumulative log odds ratios for the population of strata. In each stratum the model can also predict the cumulative log odds ratio by shrinking the sample cumulative log odds ratios toward the mean. This is especially useful when the sample size in a stratum is small and ordinary sample log odds ratios have large standard errors or are even infinite. Even when the strata are not a random sample or not even a sample and a random effects approach is not as natural, the model is beneficial for these purposes.

Table 10.5, analyzed by Hartzel et al. (2001a), is an example of this type. This table shows results for eight centers from a double-blind, parallel-group clinical study. The study was designed to compare an active drug with a placebo in the treatment of patients suffering from asthma. Patients were randomly assigned to the treatments. At the end of the study, investigators described their perception of the patient's change in condition, using the ordinal scale (much better, better, unchanged or worse).

We will use random effects models to compare the treatments while simultaneously investigating potential treatment × center interaction and modeling the association variability among centers. In doing so, even though the clinics were not randomly chosen, the assumption of a random clinic effect yields statistical inferences that better capture the variability inherent in this setting than when clinic effects are considered fixed. In addition, the random effects approach more

TABLE 10.5. Clinical Trial Relating Treatment to Response for Eight Centers

		Response		
Center	Treatment	Much Better	Better	Unchanged or Worse
1	Drug	13	7	6
	Placebo	1	1	10
2	Drug	2	5	10
	Placebo	2	2	1
3	Drug	11	23	7
	Placebo	2	8	2
4	Drug	7	11	8
	Placebo	0	3	2
5	Drug	15	3	5
	Placebo	1	1	5
6	Drug	13	5	5
	Placebo	4	0	1
7	Drug	7	4	13
	Placebo	1	1	11
8	Drug	15	9	2
	Placebo	3	2	2

Source: Data supplied by I. Liu.

naturally directs the inferences toward the true population of interest (i.e., for all such centers) rather than only these eight centers. The fixed effects models have the limitation that their inferences, strictly speaking, apply only to those centers. Ideally, for random effects modeling, we would prefer to have more than the eight strata in Table 10.5. Keeping in mind the limitations of a small number of (non-randomly chosen) centers and sparse data, we use these data to illustrate ordinal GLMMs having both a random intercept and a random treatment effect.

Let Y_{it} denote observation t in center i, and let x_{it} denote a treatment indicator variable (drug = 1, placebo = 0). The cumulative logit random-intercept model with proportional odds structure is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta x_{it}, \qquad j = 1, 2. \qquad (10.5)$$

where $\{u_i\}$ are independent $N(0, \sigma_u^2)$ variates. This model assumes that each center has the same cumulative log odds ratio β. A more general model that permits heterogeneity in the cumulative log odds ratios is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + (\beta + v_i)x_{it}, \qquad j = 1, 2, \qquad (10.6)$$

where $\{(u_i, v_i)\}$ are independent bivariate normal random effects having means $(0, 0)$, standard deviations (σ_u, σ_v), and correlation ρ. Parameters of main interest are the mean β and the standard deviation σ_v of the center-specific cumulative log

odds ratios $\{\beta + v_i\}$. We could also consider a more complex cutpoint structure that replaces $u_i + \alpha_j$ by a vector of random effects $\{u_{ij}, j = 1, \ldots, c - 1\}$. Although such a model fits better, it can present computational problems[3] for sparse data such as in Table 10.5. In addition, this more complex modeling usually has little impact on the question of main interest, namely, on estimating β and σ_v.

The homogeneity cumulative logit model (10.5), which has only a random center intercept, has $\hat{\beta} = 0.95$ and SE = 0.28. Model (10.6), which allows heterogeneity in the treatment effects among centers, has $\hat{\beta} = 0.92$ with SE = 0.53. Although the treatment estimate is similar for this more complex model, the SE is much larger. This larger SE results from the extra source of variability represented by $\{v_i\}$ and described by $\hat{\sigma}_v = 1.22$. That is, model (10.6) predicts that cumulative log odds ratios vary among centers with a mean of 0.92 and a standard deviation of 1.22. The SE of $\hat{\beta}$ in the heterogeneity model is similar to that of the corresponding $\hat{\beta}$ in a homogeneity model only when $\hat{\sigma}_v$ is close to 0.

We can test the hypothesis of homogeneity of the cumulative odds ratios across centers ($H_0: \sigma_v = 0$) by comparing these two models. The likelihood-ratio test statistic is the difference in the maximized 2(log-likelihood) values, which equals 5.9. Under H_0, this statistic has asymptotic distribution that is an equal mixture of two chi-squared distributions, with df = 1 and with df = 2. The P-value is 0.03, the average of the tail probabilities above 5.9 for χ_1^2 and χ_2^2 variates. There is considerable evidence of heterogeneity. Recognizing the heterogeneity, we must be content with a less precise estimate of the overall association level.

Table 10.6 shows the predicted values for the cumulative log odds ratios according to model (10.6), based on predicting the value of v_i and finding $\hat{\beta} + \hat{v}_i$. The predicted value $\hat{v}_i$ equals the estimate of its expected value, given the data. (The standard errors provided in the table were found using a Laplace approximation to the conditional mean-squared error of prediction.) The random effects estimates "borrow from the whole" using data from all the centers to estimate the cumulative log odds ratio in each center. Because of this, they show considerable shrinkage toward the mean compared to the estimates from a fixed effects model that allows heterogeneous effects, which uses the data in a center alone to estimate the cumulative log odds ratio in that center. The table also shows these estimates. For instance, the negative estimates of -1.62 and -1.06 shrink to -0.62 and -0.10. When data sets have small sample sizes per stratum, shrinkage of effect estimates is highly appealing for models permitting heterogeneity, since the stratum-specific sample estimates are then likely to exhibit more variability than the true parameters. In particular, a stratum-specific estimate is infinite when none of the sample pairs of observations in the stratum are concordant or none are discordant.

Next, we consider the significance of the treatment effect, by testing $H_0 : \beta = 0$. We begin with the homogeneous effect model, but only for illustrative purposes, since it fits poorly. For the model (10.5) with random center effects, the likelihood-ratio statistic equals 12.0 ($P < 0.001$). This test, coupled with the positive sign for $\hat{\beta}$, provides strong evidence that the response tends to be better with drug than with

---

[3]See www.stat.ufl.edu/~aa/ordinal/data.html for SAS NLMIXED code.

TABLE 10.6. Summary of Center-Specific Cumulative Log Odds Ratio Estimates and Standard Errors for Treatment Effects with Fixed and Random Effects Heterogeneity Models Applied to Table 10.5

Effect	Fixed Effects Model		Random Effects Model	
	Estimate	*SE*	Estimate	*SE*
Center 1	3.03	0.87	2.35	0.75
Center 2	−1.62	0.95	−0.62	0.92
Center 3	0.20	0.55	0.32	0.52
Center 4	0.71	0.85	0.76	0.72
Center 5	2.84	0.95	2.11	0.83
Center 6	−1.06	1.21	−0.10	0.94
Center 7	1.76	0.87	1.53	0.73
Center 8	0.83	0.82	0.84	0.73

placebo. However, we've seen that the homogeneity assumption is unrealistic. In the heterogeneous association model (10.6), the likelihood-ratio statistic for testing $H_0 : \beta = 0$ that the mean of the cumulative log odds ratios is zero equals 2.5 with df = 1 (*P*-value = 0.11 for $H_a : \beta \neq 0$). The evidence of a treatment effect is considerably weaker, and that effect is then a *mean* effect rather than a common effect for each stratum. In using more realistic models permitting heterogeneity, it can be more difficult to establish significance of effects because of the extra variability inherent in the model.

Similar substantive results occur for corresponding models using adjacent-categories logits. With a heterogeneity model, $\hat{\beta} = 0.63$ with a standard error of 0.34. Again the standard error is considerably larger than for a homogeneity model, for which $\hat{\beta} = 0.65$ with a standard error of 0.19. The variability among $\{v_i\}$ is described by $\hat{\sigma}_v = 0.77$.

These models summarized between–strata heterogeneity but have a single parameter describing association within each stratum. Although it is also unrealistic to think that all odds ratios within a stratum are truly exactly equal, in practice it is often sufficient to summarize the overall association within a stratum and describe the variability in that overall association across strata. An alternative approach models the heterogeneity both within and between strata. See Hartzel et al. (2001a) and Coull and Agresti (2003).

10.3.5 Example: Toxicity Study Using Continuation-Ratio Logits

For continuation-ratio logit models with ordinal responses, the logits refer to independent binomial variates (Section 4.2). Thus, binary logistic random effects models apply to clustered ordinal responses using continuation-ratio logits (e.g., Ten Have and Uttal 1994). For observation t in cluster i, let

$$\omega_{ij} = P(Y_{it} = j \mid Y_{it} \geq j; \; u_{ij}).$$

Let n_{ij} be the number of subjects in cluster i that make response j. Let $n_i = \sum_{j=1}^{c} n_{ij}$. For a given cluster in a continuation-ratio logit model, treating $n_{i1}, \ldots, n_{i,c-1}$ as multinomial is equivalent to treating the separate counts as a sequential set of independent binomial variates, where n_{ij} is binomial for $(n_i - \sum_{h<j} n_{ih})$ trials with parameter ω_{ij}, $j = 1, \ldots, c - 1$.

We illustrate with a developmental toxicity study conducted under the U.S. National Toxicology Program. This study examined the developmental effects of ethylene glycol (EG) by administering one of four dosages (0, 0.75, 1.50, 3.00 g/kg) to pregnant rodents. The four dose groups had (25, 24, 22, 23) pregnant rodents. The clusters are litters of mice. The three possible outcomes are dead/resorption, malformation, and normal. Table 10.7 shows the data. The continuation-ratio logit is natural here because categories are hierarchically related, in the sense that an animal must survive before a malformation can take place. Coull and Agresti (2000) presented the following analyses.

For litter i in dose group d, let logit$[\omega_{i(d)1}]$ be the continuation-ratio logit for the probability of death, and let logit$[\omega_{i(d)2}]$ be the continuation-ratio logit for the conditional probability of malformation, given survival. (The notation $i(d)$ represents litter i nested within dose d.) Let x_d be the dosage for group d. We account for the litter effect using litter-specific random effects $\mathbf{u}_{i(d)} = (u_{i(d)1}, u_{i(d)2})$ sampled from a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_d that may vary according to the dosage level x_d. This bivariate random effect allows for differing amounts of overdispersion for the probability of death and for the probability of malformation, given survival. A model also permitting different fixed effects for each is

$$\text{logit}[\omega_{i(d)j}] = u_{i(d)j} + \alpha_j + \beta_j x_d. \tag{10.7}$$

TABLE 10.7. Response Counts for 94 Litters of Mice on Numbers (Dead, Malformed, Normal) for Fetuses in the Litters

Dose = 0.00 g/kg	Dose = 0.75 g/kg	Dose = 1.50 g/kg	Dose = 3.00 g/kg
(1, 0, 7), (0, 0, 14)	(0, 3, 7), (1, 3, 11)	(0, 8, 2), (0, 6, 5)	(0, 4, 3), (1, 9, 1)
(0, 0, 13), (0, 0, 10)	(0, 2, 9), (0, 0, 12)	(0, 5, 7), (0, 11, 2)	(0, 4, 8), (1, 11, 0)
(0, 1, 15), (1, 0, 14)	(0, 1, 11), (0, 3, 10)	(1, 6, 3), (0, 7, 6)	(0, 7, 3), (0, 9, 1)
(1, 0, 10), (0, 0, 12)	(0, 0, 15), (0, 0, 11)	(0, 0, 1), (0, 3, 8)	(0, 3, 1), (0, 7, 0)
(0, 0, 11), (0, 0, 8)	(2, 0, 8), (0, 1, 10)	(0, 8, 3), (0, 2, 12)	(0, 1, 3), (0, 12, 0)
(1, 0, 6), (0, 0, 15)	(0, 0, 10), (0, 1, 13)	(0, 1, 12), (0, 10, 5)	(2, 12, 0), (0, 11, 3)
(0, 0, 12), (0, 0, 12)	(0, 1, 9), (0, 0, 14)	(0, 5, 6), (0, 1, 11)	(0, 5, 6), (0, 4, 8)
(0, 0, 13), (0, 0, 10)	(1, 1, 11), (0, 1, 9)	(0, 3, 10), (0, 0, 13)	(0, 5, 7), (2, 3, 9)
(0, 0, 10), (1, 0, 11)	(0, 1, 10), (0, 0, 15)	(0, 6, 1), (0, 2, 6)	(0, 9, 1), (0, 0, 9)
(0, 0, 12), (0, 0, 13)	(0, 0, 15), (0, 3, 10)	(0, 1, 2), (0, 0, 7)	(0, 5, 4), (0, 2, 5)
(1, 0, 14), (0, 0, 13)	(0, 2, 5), (0, 1, 11)	(0, 4, 6), (0, 0, 12)	(1, 3, 9), (0, 2, 5)
(0, 0, 13), (1, 0, 14)	(0, 1, 6), (1, 1, 8)		(0, 1, 11)
(0, 0, 14)			

Source: Study described in article by C. J. Price, C. A Kimmel, R. W. Tyl, and M. C. Marr, *Toxicol. Appl. Pharmacol.* **81**, 113–127 (1985).

TABLE 10.8. Comparisons of Log Likelihoods for Multivariate Random Effects Models for the Developmental Toxicity Study Summarized by Table 10.7

Model	Number of Parameters	Change in Parameters	Change in Log Likelihood
Dose-specific Σ_i	16		
Σ_i, common α, β	14	2	28.4
Common Σ	7	9	7.4
Common Σ, $\rho = 0$	6	10	7.4
Univariate σ^2	5	11	16.7

Table 10.8 reports the change in the maximized log likelihood from fitting four special cases of this model:

- Common intercept and slope for the two logits $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$.
- Common covariance matrix for the four doses $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$.
- Common covariance matrix and uncorrelated random effects.
- Univariate common variance components across dose: $u_{i(d)1} = u_{i(d)2}$ and $\sigma_d = \sigma_u$.

Tests of the first three special cases against the general model (10.7) can use ordinary likelihood-ratio tests. It seems adequate to use the simpler model having uncorrelated random effects with homogeneous covariance structure (i.e., the fourth model listed in Table 10.8), since the likelihood-ratio statistic comparing this to model (10.7) equals $2(7.4) = 14.8$ (df $= 10$). The model provides a separate model for each conditional binomial outcome, specifying that the proportion of dead pups and the proportion of malformed pups (given survival) are independent, both within litter and marginally.

The univariate model in Table 10.8 is the special case of the third model listed in which the variances are common for the two logits and the random effects are perfectly correlated. Hence, it reduces to a univariate random effects model. According to standard model selection criteria such as the AIC, the univariate model is inadequate.

The ML estimated effects for the separate model for each conditional binomial outcome are $\hat{\beta}_1 = 0.08$ (SE $= 0.21$), $\hat{\beta}_2 = 1.79$ (SE $= 0.22$). For a given cluster, there is no evidence of a dose effect on the death rate, but the estimated odds of malformation, given survival, multiply by $\exp(1.79) = 6.0$ for every additional g/kg of ethylene glycol. The variance component estimates suggest a stronger litter effect for the malformation outcome given survival ($\hat{\sigma}_2 = 1.6$) than for death ($\hat{\sigma}_1 = 0.5$).

10.3.6 More Complex Models with Multiple Random Effects

Cumulative logit model (10.4) has the proportional odds structure

$$\text{logit}[P(Y_{it} \leq j)] = \alpha_j + \mathbf{u}_i'\mathbf{w}_{it} + \boldsymbol{\beta}'\mathbf{x}_{it}, \qquad j = 1, \ldots, c - 1, \qquad (10.8)$$

by which the effects are the same for each cumulative probability. More general models permit fixed effects $\boldsymbol{\beta}_j$ that vary for different cumulative logits for at least some predictors (i.e., partial proportional odds, as in Section 3.6.4). However, such models have a structural problem whereby cumulative probabilities may be misordered at some predictor settings. See Hedeker and Mermelstein (1998) and Hedeker and Gibbons (2006, Secs. 10.2.1, 10.2.4).

The proportional odds structure fails when different groups have different dispersion. An alternative approach generalizes the model introduced in Section 5.4, which includes dispersion effects. The generalized model has the form

$$\text{logit}[P(Y_{it} \leq j)] = \frac{\alpha_j - \mathbf{u}_i'\mathbf{w}_{it} - \boldsymbol{\beta}'\mathbf{x}_{it}}{\exp(\boldsymbol{\gamma}'\mathbf{x}_{it})}.$$

See Ishwaran and Gatsonis (2000) and Hedeker et al. (2006) for examples.

In model (10.8), each logit for cluster i has the same random effect $\mathbf{u}_i$. Such simplifications result naturally from underlying latent variable models for which the latent response variable has a logistic distribution. Tutz and Hennevogl (1996) considered a more general model that allowed a different random effect $\mathbf{u}_{ij}$ for each cumulative logit. This is the natural approach for a baseline-category logit model with random effects for a nominal response variable. Estimation in this extended model is more complicated, because the intercepts must be reparameterized to ensure that their ordering is not violated.

Alternative ordinal models use other ordinal logits or link functions. The complementary log-log link is useful for survival data (Hedeker and Gibbons 2006, Sec. 10.2.3; Rabe-Hesketh and Skrondal 2008, Secs. 8.6–8.8). When the adjacent-categories logit (ACL) model has the same predictor form as model (10.8), $\boldsymbol{\beta}$ has log odds interpretations for all pairs of adjacent categories. Since intercepts in the ACL model are unordered, an extended model permitting different random effects for each logit does not require reparameterization and has the form

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = j + 1)} = \alpha_j + \mathbf{u}_{ij}'\mathbf{w}_{it} + \boldsymbol{\beta}'\mathbf{x}_{it}, \qquad j = 1, \ldots, c - 1. \qquad (10.9)$$

The ACL model is equivalent to a baseline-category logit (BCL) model with an adjusted model matrix. The random effect $\mathbf{u}_{ij}$ in the BCL model then corresponds to $\mathbf{u}_{ij} + \cdots + \mathbf{u}_{i,c-1}$ from the ACL model. Thus, a simple covariance structure for the random effects in the ACL model implies a certain pattern for the covariance matrix for the random effects in the baseline-category logit model. An advantage of such a general model is that it may fit better. A disadvantage is that having a larger vector of random effects makes model fitting more challenging. When the main interest is in estimating the fixed effects, the simpler model is usually adequate.

10.4 MULTILEVEL (HIERARCHICAL) ORDINAL MODELS

Hierarchical models describe observations that have a nested nature: Units at one level are contained within units of another level. Such data are common in certain

application areas, such as in educational studies. Models that take into account the hierarchical nature of the observations are called *multilevel models*. Random effects can enter the model at each level of the hierarchy.

Simple random effects models having a random intercept can be regarded as multilevel models. The first level consists of the observations that are nested within the clusters. The second level consists of the clusters themselves, represented by the random effects. In a three-level model, the second level would itself be nested within a third level, such as when multiple observations (level 1) on a student (level 2) occur within different schools (level 3). The following example of a three-level model is based on a data set analyzed by Rampichini et al. (2004).

10.4.1 Example: Student Evaluations of University Courses

Students at the University of Florence (Italy) evaluate several aspects of their courses, such as the clarity of the teaching, the clarity of the exam rules, and the teacher's ability to answer student questions. Each rating uses a four-category scale: decidedly negative, more negative than positive, more positive than negative, decidedly positive. We refer to the different aspects of the course that are evaluated as *items*. A model for analyzing factors that affect the student evaluation on a particular item might use explanatory variables that include (a) characteristics of the student, such as GPA, whether a full-time student, gender, level of student interest in the course, whether the student's previous knowledge was sufficient for meeting the demands of the course, and the student's level of expectations for the course; and (b) characteristics of the course, such as the subject matter, whether the course is required or an elective, and the amount of required reading material.

Just as responses for two items by the same student might tend to be more alike than responses for those items by different students, so might responses on a particular item by two students in the same course tend to be more alike than responses on that item by students from different courses. Student and course might be treated as random effects, with different ones referring to different levels of the model. For example, a model might have the several evaluations of a course by a student at level 1, a random effect for each student at level 2, and a random effect for the course at level 3. A multilevel cumulative link model then has a random effect for each student in a course, a random effect for each course, and fixed effects for a set of explanatory variables.

Let $y_{i(k)t}$ denote the ordinal evaluation response for student i in course k for item t in the battery of items evaluated. A linear model for an unobserved latent response variable $y^*_{i(k)t}$ naturally induces a cumulative link model, as observed in Section 3.3.2. Let $\mathbf{x}_{i(k)t}$ denote the values of predictor variables for this evaluation. Suppose that

$$y^*_{i(k)t} = u_k + v_{i(k)} + \boldsymbol{\beta}'\mathbf{x}_{i(k)t} + \epsilon_{i(k)t},$$

where $\epsilon_{i(k)t}$ has a standardized distribution such as the standard normal. We observe $y_{i(k)t}$ in outcome category j if $y^*_{i(k)t}$ falls between α_{j-1} and α_j. The model for the observed response $y_{i(k)t}$ then has the form

$$\text{link}\,[P(y_{i(k)t} \le j)] = u_k + v_{i(k)} + \alpha_j + \boldsymbol{\beta}'\mathbf{x}_{i(k)t}.$$

This is a cumulative probit model when we assume a normal distribution for $\epsilon_{i(k)t}$ and a cumulative logit model when we assume a logistic distribution.

Here the explanatory variables **x** would include one that identifies the item evaluated. The random effects u_k for courses and $v_{i(k)}$ for students within courses are independent with different variance components. The student random effects $\{v_{i(k)}\}$ account for variability among students in characteristics not measured by **x**. When they have a relatively large variance component, there is a strong correlation among the various ratings by a student (recall Section 10.1.3). The course random effects $\{u_k\}$ account for variability among courses. Predicted values of these describe the effect of the course on the student evaluations, conditional on the explanatory variables and the student random effect. They serve as adjusted indicators of the course quality.

For this model, two ratings of the same course are correlated, but conditional on the explanatory variables, two ratings of different courses are uncorrelated. In practice, the same student may rate more than one course. This requires more complex modeling, because the student is then partially cross-classified with the course rather than nested within a single course. See Rampichini et al. (2004) and Skrondal and Rabe-Hesketh (2004, pp. 60–63, 321–348) for discussion, references, and examples.

10.4.2 Example: Effectiveness of Sex Education

In this section we describe an example analyzed in detail in Skrondal and Rabe-Hesketh (2004, Sec. 10.2), based on a study that evaluated a sex education program for 15- and 16-year-old students in Norway. Schools participating in the study were randomly assigned to administer or not administer the sex education course. We refer to the two groups of schools as the treatment group and the control group. Students were observed at the time of randomization and then six months and 18 months after the randomization, in terms of their responses to the statement, "If my partner and I were about to have intercourse without either of us having mentioned contraception, I could easily get out a condom (if I had one with me)." The ordered response categories were (not at all true, slightly true, somewhat true, mostly true, completely true).

A model can contain random effects for the students to account for the likely positive association for repeated observations on the same student. It can also contain random effects for the schools to account for observations from students in the same school tending to be more alike that observations from students in different schools. Let $y_{i(k)t}$ denote the response for student i in school k at time t following the randomization, measured in six-month multiples ($t = 0$ for initial, $t = 1$ for six months, $t = 3$ for 18 months). Let x be an indicator of whether the student was in the treatment group of schools ($1 =$ yes, $0 =$ no). The model

$$\text{logit}[P(y_{i(k)t} \leq j)] = u_k + v_{i(k)} + \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t \times x)$$

allows the time trend to differ for the treatment and control groups. This model, with two random intercepts, has the same form as described in the preceding

example. Skrondal and Rabe-Hesketh reported that the school random effect u_k had estimated variance of 0 and could be dropped from the model. By contrast, the within-subject associations between pairs of times were reflected by a variance estimate of 2.03 (SE = 0.31) for the subject random effect $v_{i(k)}$.

For this model or the simpler one without the school random effect, the fixed effects estimates were $\hat{\beta}_1 = 0.13$ (SE = 0.06), $\hat{\beta}_2 = 0.02$ (SE = 0.19), and $\hat{\beta}_3 = -0.17$ (SE = 0.09). The $\hat{\beta}_2$ value estimates the treatment effect initially, and this was essentially 0 as expected by the randomization. The time trend was estimated to be $\hat{\beta}_1 = 0.13$ (perhaps surprisingly) for the control group and $\hat{\beta}_1 + \hat{\beta}_3 = -0.04$ for the treatment group. The difference $-\hat{\beta}_3 = 0.17$ and the corresponding cumulative odds ratio of $\exp(0.17) = 1.19$ indicates that after six months the change in the odds of a relatively high response instead of one below that is 19% higher in the treatment group than in the control group. But this effect is only marginally statistically significant.

At each time, students also responded to two other statements using the same response scale: "If my partner and I were about to have intercourse without either of us having mentioned contraception, I could easily tell him/her that I didn't have any contraception" and "If my partner and I were about to have intercourse without either of us having mentioned contraception, I could easily ask him/her if he/she had any contraception." One can build a model that considers all three items simultaneously. Let $y_{i(k)st}$ denote the response for student i in school k to item s at time t after the randomization. If responses merely tend to shift in location for the various items, the model

$$\text{logit}[P(y_{i(k)st} \leq j)] = u_i + v_{i(k)} + \alpha_j + \delta_s + \beta_1 t + \beta_2 x_1 + \beta_3(t \times x_1)$$

with a constraint such as $\delta_1 = 0$ may be adequate.

This model could be generalized in various ways. To allow the item effects to vary by time, we could add an item × time interaction term. More complex random effect structure can also be useful: for instance to allow the associations among the repeated responses for the various times to differ according to the item. In such a case, it is still often sensible to expect that a student who has a relatively high response on one item will tend to have a relatively high response on another item. Then we could allow the random effect for each subject to be scaled larger or smaller as we move from item to item. One way to do this is to replace the subject random effect u_i in the model by a term $\lambda_s u_i$, where u_i is a standard normal random variable. Then the standard deviation of the random effects is λ_s for item s. More generally, we could let the subject random effects vary by the item or vary by the occasion. See Skrondal and Rabe-Hesketh (2004) for details.

10.5 COMPARING RANDOM EFFECTS MODELS AND MARGINAL MODELS

In Chapters 9 and 10 we have presented three types of models for clustered, ordinal data: (1) Marginal models describe each Y_t separately in terms of explanatory

variables; (2) transitional models describe each Y_t in terms of previous response observations such as y_{t-1} as well as the explanatory variables; and (3) random effects models describe each Y_{it} in terms of explanatory variables, while sharing a random effect u_i by the observations within cluster i to induce a joint distribution among the clustered observations.

If we want to use past observations to predict future observations, thus describing effects of explanatory variables conditional on those past observations, it is natural to use transitional models. The choice between marginal models and random effects models is not as clear-cut, because they both deal with using each observation in a cluster as a response variable. In this section we describe differences in interpretations for these two types of models and discuss factors that can influence the choice of the type of model.

10.5.1 Differing Effects in Random Effects Models and Marginal Models

As explained in Sections 8.2.6 and 10.1.4, effects of explanatory variables have different interpretations in random effects models than in marginal models. The parameters in random effects models have conditional, cluster-specific intepretations, given the random effect. By contrast, effects in marginal models are averaged over clusters (i.e., population averaged).

The two types of effects not only have different interpretations but can have quite different sizes. In the arthritis study, the parameter estimates for the random effects model of Section 10.2.2 were about 50% larger than the estimates for the marginal model of Section 9.2.3 (see Table 10.2). One way to understand why this happens is to recall from Section 8.2.6 that a cluster-specific model applies naturally to probabilities for the data as displayed in a separate partial table for each cluster. By contrast, a marginal model applies naturally to the probabilities for the data collapsed over these partial tables. Basic results in contingency table analysis, such as Simpson's paradox, tell us that marginal associations can be quite different from the conditional associations.

Another way to understand the distinction is to note that the difference in sizes of effects is caused by the link function being nonlinear. Figure 10.1 illustrates this for the first cumulative probability ($j = 1$) at a fixed value of t for a cumulative logit random intercept model. For a single quantitative explanatory variable x, the figure shows cluster-specific cumulative logistic curves for $P(Y_{it} = 1 \mid u_i)$ for several clusters. The rate of increase in the curves reflects the size of the coefficient β of x. In this figure, the considerable spread among the curves reflects substantial heterogeneity among the random effects $\{u_i\}$, corresponding to a relatively large σ_u. At any fixed x-value, variability occurs in values of $P(Y_{it} = 1 \mid u_i)$ for different i, and the average of these values is the marginal $P(Y_t = 1)$. These averages for various x-values yield the superimposed dashed curve that represents the marginal model. For that marginal (dashed) curve, the effect is considerably weaker than for each separate subject-specific curve.

In summary, population-averaged effects in marginal models are smaller in magnitude than cluster-specific effects in random effects models. The difference
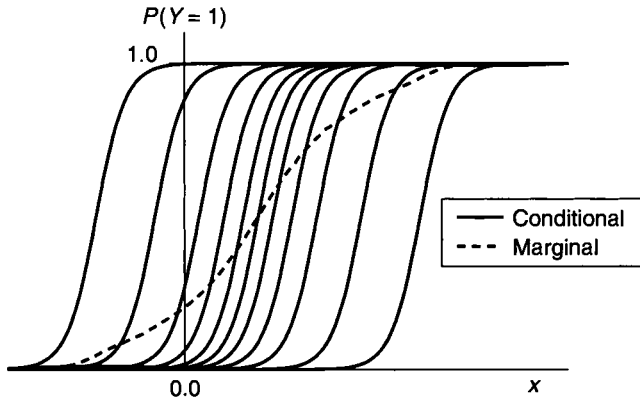
Figure 10.1. Logistic random-intercept model, showing the conditional (cluster-specific) curves and the marginal (population-averaged) curve averaging over these.

between the two effects increases as the cluster-specific curves are more spread out, that is, as σ_u increases. This is caused by the nonlinearity of the link function. By contrast, suppose that the curves were replaced by straight lines, corresponding to using the identity link function. Then the line connecting the average values of $P(Y_{it} = 1 \mid u_i)$ would have the same slope as that of each subject-specific line.

In Section 10.1.3 we observed that as the random effects variability σ_u increases, so does the within-cluster correlation. So the stronger the association among responses within a cluster, the greater the difference between the parameter values in a random effects model and in the corresponding marginal model. To illustrate, relatively large heterogeneity was estimated for the cumulative logit model for the arthritis data in Section 10.2.2, with $\hat{\sigma}_u = 1.92$. Because of this, the estimates of the effects were considerably larger than those obtained for the corresponding marginal model in Section 9.2.3.

Another relevant point here is that when a cumulative link random effects model holds, when we integrate out the random effects to obtain the corresponding marginal model, that marginal model does not exactly have the cumulative link model form, except when $\sigma_u = 0$. An exception occurs with the probit link. In that case, the cumulative probit random intercept model *does* imply a marginal model of cumulative probit form. Then, an effect in the probit random effects model equals the corresponding effect in the probit marginal model multiplied by $(1 + \sigma_u^2)^{1/2}$. When $\sigma_u = (0, 1, 2, 3)$, the ratio of the effect in the random effects model to the effect in the marginal model is $(1.0, 1.4, 2.2, 3.2)$.

When the cumulative logit random intercept model holds, the marginal model does not have cumulative logit form, but a cumulative logit model approximates the actual marginal model well. The effect in the random effects model equals the corresponding effect in the marginal model multiplied by approximately $(1 + 0.346\sigma_u^2)^{1/2}$. [The 0.346 multiple occurs because the standard logistic cdf at a point z is well approximated by the standard normal cdf at the point $(16\sqrt{3}/15\pi)z =$

0.588z, and $(0.588)^2 = 0.346$.] When $\sigma_u = (0, 1, 2, 3, 4)$, the ratio of the effect in the random effects model to the effect in the marginal model is about $(1.0, 1.2, 1.5, 2.0, 2.6)$.

10.5.2 Example: Repeated Responses on Similar Survey Items

Table 8.11 in Exercise 8.5 showed GSS data regarding government spending on (1) assistance to big cities, (2) law enforcement, (3) health, and (4) the environment, using the response scale (too little, about right, too much). Consider the random effects model

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_t x_{it}, \qquad j = 1, 2, \quad t = 1, 2, 3, 4,$$

where x_{it} is an indicator variable that equals 1 for item t and equals 0 otherwise. With the constraint $\beta_4 = 0$, Table 10.9 shows the ML estimates for β_1, β_2, β_3, and σ_u. By contrast, that table also shows the estimates for the corresponding marginal model, based on the GEE approach with independence working correlation. From the discussion above, the random effects model estimates should equal approximately the marginal model estimates multiplied by $(1 + 0.346\hat{\sigma}_u^2)^{1/2} = 1.10$. The actual ratios of random effects model estimates to marginal model estimates are close to this, being 1.11 for $\hat{\beta}_1$, 1.10 for $\hat{\beta}_2$, and 1.16 for $\hat{\beta}_3$. The SE values have similar relative sizes, and the two approaches provide similar substantive conclusions.

Let's compare effect interpretations, illustrating using $\hat{\beta}_1$. For the marginal model, $\hat{\beta}_1 = -2.338$ means that *for the population marginal distributions*, the estimated odds of response "too little" rather than "about right" or "too much" for spending on the environment were $\exp(2.338) = 10.4$ times the corresponding estimated odds for spending on cities. By contrast, for the random effects model, $\hat{\beta}_1 = -2.586$ means that *for a given subject*, the estimated odds of response "too little" rather than "about right" or "too much" for spending on the environment were $\exp(2.586) = 13.3$ times the corresponding estimated odds for spending on cities.

Next let's compare probability prediction, illustrating using the first category ("too little"). Using predicted random effect values for the random effects model, given the data we can make subject-specific predictions. For example, each subject in the first cell (response in category 1 for each item) has predicted random effect

TABLE 10.9. Results for Cumulative Logit Models Fitted to Table 8.11, with SE Values in Parentheses

Estimate	Random Effects Model	Marginal Model (GEE)
$\hat{\beta}_1$	−2.586 (0.137)	−2.338 (0.121)
$\hat{\beta}_2$	−0.510 (0.128)	−0.465 (0.119)
$\hat{\beta}_3$	−0.088 (0.133)	−0.076 (0.116)
$\hat{\sigma}_u$	0.78 (0.08)	

$\hat{u}_i = 0.718$. The ML estimate of the first cutpoint parameter is $\hat{\alpha}_1 = 1.114$, so an estimated probability of category 1 for the response on cities is

$$\frac{\exp(\hat{u}_i + \hat{\alpha}_1 + \hat{\beta}_1)}{1 + \exp(\hat{u}_i + \hat{\alpha}_1 + \hat{\beta}_1)} = \frac{\exp(0.718 + 1.114 - 2.586)}{1 + \exp(0.718 + 1.114 - 2.586)} = 0.32.$$

By comparison, each subject in the last cell (response in category 3 for each item) has $\hat{u}_i = -1.607$, so an estimated probability of category 1 for the response on cities is then only

$$\frac{\exp(\hat{u}_i + \hat{\alpha}_1 + \hat{\beta}_1)}{1 + \exp(\hat{u}_i + \hat{\alpha}_1 + \hat{\beta}_1)} = \frac{\exp(-1.607 + 1.114 - 2.586)}{1 + \exp(-1.607 + 1.114 - 2.586)} = 0.04.$$

The marginal model does not generate such subject-specific estimates. For that model, $\hat{\alpha}_1 = 0.997$, so an estimated population-averaged probability of category 1 for the response on cities is

$$\frac{\exp(\hat{\alpha}_1 + \hat{\beta}_1)}{1 + \exp(\hat{\alpha}_1 + \hat{\beta}_1)} = \frac{\exp(0.997 - 2.338)}{1 + \exp(0.997 - 2.338)} = 0.21.$$

10.5.3 Choosing Between Random Effects Models and Marginal Models

In Chapters 9 and 10 we have surveyed the use of marginal models and random effects models for clustered ordinal data. Agresti and Natarajan (2001) provided an earlier survey. What are the issues that affect the choice of one type of model over the other?

We have seen that effect parameters are larger in random effects models than in marginal models, and more so as variance components increase. Usually, though, the significance of an effect is similar for the two model types. If one effect seems more important than another in a random effects model, the same is usually true with a marginal model. The choice of the model is usually not crucial to inferential conclusions. This statement requires a caveat, however, since sizes of effects in marginal models depend on the degree of heterogeneity in random effects models. In comparing effects for two groups or two variables that have quite different variance components, relative sizes of effects will differ for marginal and random effects models. The attenuation from the conditional to the marginal effect will tend to be greater for the group having the larger variance component. This remark is especially relevant when we allow different variability of random effects for two groups, such as mentioned near the end of the example in Section 10.3.3.

Some statisticians prefer models with random effects over marginal models, because they more fully describe the structure of the data. The model provides a joint distribution for all the observations in a cluster and hence a likelihood function that can be used in the same way as with other statistical models. However, many statisticians believe that both model types are useful and that the choice of model

should depend on the purpose of the study. Different questions can suggest different models for the same data.

Latent variable constructions used to motivate model forms such as the cumulative probit and cumulative logit usually apply more naturally at the cluster level than the marginal level. Given a random effects model, we can in principle generate the estimated marginal distributions, although this may require extra work not readily done by standard software. That is, a random effects model implies a marginal model. For example, with the cumulative probit link the marginal model also satisfies a cumulative probit model, but with weaker effects. By contrast, a marginal model does not itself imply a random effects model, although it does implicitly determine the form of the fixed portion of the linear predictor in such a model, as defined by the integral equation linking the marginal and conditional cumulative probabilities (Heagerty and Zeger 2000b). In this sense, a random effects model contains more information than is contained in a marginal model.

The random effects modeling approach is preferable if you want to fully model the joint distribution. That approach is also preferable if you want to estimate cluster-specific effects or estimate their variability, or if you want to specify a mechanism that could generate nonnegative association among clustered observations. Because a marginal model does not explicitly include random effects, it does not allow estimation of cluster-specific effects or probabilities. Some methodologists use random effects models whenever the main goal is to learn about within-cluster effects. In crossover studies such as the one analyzed in Section 10.2.1, the random effects model is appropriate to obtain an estimate of the effects comparing treatments that is "within-subject," describing differences among treatments at the subject level.

By contrast, if the main focus is on comparing groups that are independent samples, effects of interest are of between-cluster rather than within-cluster type. It is then often adequate to estimate effects with a marginal model. For example, if after a period of time we mainly want to compare the relative frequency of the best response category for those taking a new drug and for those taking a standard drug, a marginal model is adequate. In many surveys or epidemiological studies, the principal goal is to compare the relative frequency of occurrence of some outcome for different groups in a population. Then quantities of primary interest include between-group odds ratios comparing marginal probabilities for the various groups. When marginal effects are the main focus, it is simpler to model the margins directly and avoid trying to model aspects of the joint distribution that are not of particular interest in the study. One can then parameterize the model so that regression parameters have a direct marginal interpretation. Developing a more detailed model of the joint distribution that generates those margins, as a random effects model does, provides greater opportunity for misspecification. For example, with longitudinal data the assumption that observations are independent, given the random effect, need not be realistic. Or, the shape of the distribution of the random effects may be far from normal, or the variance of the distribution of the random effects may be quite different for some groups than for others. Having severe model misspecification can result in badly biased estimates of effects.

In summary, for studies that focus on the individual or the cluster, random effects models are natural, whereas for studies that focus on the population, marginal models are natural. For example, consider applications in medical research. Marginal models are usually sufficient for public policy decisions and epidemiological purposes. However, for scientific understanding and clinical prediction, random effects models that focus on the individual as well as transitional models that use an individual's past observations to help predict the future are more informative.

Finally, for the marginal model approach, we should recall a computational issue. ML is sometimes possible (e.g., with specialized functions such as the mph.fit function in R discussed in the Appendix), as shown for the crossover study in Section 8.4.3 and the diabetic retinopathy study in Section 9.1.3. However, the GEE approach is computationally simpler, especially for large T or mixed-cluster sizes, and it is more readily available with standard software. A drawback of the GEE approach is that likelihood-based inferences are not possible because the marginal model refers only to each marginal distribution and does not specify a joint distribution of the responses. Also, with missing data, potential bias is avoided with ML inference as long as data are missing at random, whereas GEE methods have the stronger requirement that data must be missing completely at random (see Section 9.2.5).

CHAPTER NOTES

Section 10.1: Ordinal Generalized Linear Mixed Models

10.1. For factor analysis and latent class types of models for ordinal responses, see Samejima (1969), Bartholomew (1980, 1983), Croon (1990), Agresti and Lang (1993b), Qu et al. (1995), Bradlow and Zaslavsky (1999), Johnson and Albert (1999), Uebersax (1999), Anderson and Vermunt (2000), Moustaki (2000, 2003), Vermunt (2001), Quinn (2004), Todem et al. (2007), DeSantis et al. (2008), and Rabe-Hesketh and Skrondal (2008, Chap. 7).

10.2. Harville and Mee (1984) proposed a cumulative probit random effects model that utilized Taylor series approximations for intractable integrals. Jansen (1990) and Ezzet and Whitehead (1991) proposed random intercept cumulative probit and logit models, respectively, and employed quadrature techniques. Böckenholt (1999) presented a Markov cumulative probit random intercept model. Kauermann (2000) used local smoothing in a cumulative logit model for longitudinal data. More general ordinal random effects models that allowed multiple random effects were proposed by Hedeker and Gibbons (1994, 2006), who applied Gauss–Hermite quadrature within a Fisher scoring algorithm, and by Tutz and Hennevogl (1996), who used quadrature and Monte Carlo EM algorithms. Chapter 10 of Hedeker and Gibbons (2006) gave more examples of such models. For subjects measured on both ordinal and continuous response variables, Catalano (1997) assumed an underlying bivariate normal distribution in modeling clustered data. Other links have received less attention. With the complementary log-log link function, the

likelihood function for the random intercept model has closed form when the random effects distribution is the log gamma (Farewell 1982; Crouchley 1995; Ten Have 1996). Zayeri et al. (2005) applied a power transformation to the cumulative probabilities that gives the logit and complementary log-log cumulative links as special cases. Estimating the appropriate power gives information about an appropriate link function. For continuation-ratio logit models with random effects, see Ten Have and Uttal (1994), who modeled multiple discrete-time survival profiles, Coull and Agresti (2000), Dos Santos and Berridge (2000), Ten Have et al. (2000), Grilli (2005), and Rabe-Hesketh and Skrondal (2008, Chap. 8). Models for survival that use random intercepts for the subjects are referred to as *frailty models*. Xie et al. (2000) proposed random effects modeling of interval-censored ordinal data.

10.3. Hartzel et al. (2000a,b) used a semiparametric approach to fitting multinomial models with random effects. The basic models have the same multinomial form as in the fully parametric case, but a nonparametric part of the analysis estimates mass points and their probabilities for an unspecified discrete random effects distribution. This approach avoids the need for numerical integration.

Section 10.2: Examples of Ordinal Random Intercept Models

10.4. Ohman-Strickland and Lu (2003) presented power and sample size calculations for comparing two groups of subjects on an ordinal outcome in experiments where the subjects are measured before and after intervention, using ordinal random intercept models with treatment, time, and treatment-by-time interaction terms. Xiang et al. (2008) proposed influence diagnostics based on the effect on the parameter estimates of deleting clusters.

Section 10.4: Multilevel (Hierarchical) Ordinal Models

10.5. Literature on multilevel ordinal models includes Hedeker and Gibbons (1994, 2006), Hedeker and Mermelstein (1998), Qu and Tan (1998), Zaslavsky and Bradlow (1999), Fielding (1999), Ribaudo et al. (1999), Grilli and Rampichini (2002, 2003, 2007), Fielding et al. (2003), Rampichini (2004), Skrondal and Rabe-Hesketh (2004, Chap. 10), Fielding and Yang (2005), Raman and Hedeker (2005), Hedeker et al. (2006), Liu and Hedeker (2006), Plewis et al. (2006), Steele and Goldstein (2006), and Rabe-Hesketh and Skrondal (2008, Chap. 7). For Bayesian approaches, see the references in Section 11.5.1 and Note 11.4.

Section 10.5: Comparing Random Effects Models and Marginal Models

10.6. It is possible for a model to combine elements of both marginal models and random effects models. For example, for a bivariate ordinal response, Todem et al. (2007) modeled an underlying bivariate latent variable with a linear mixed model while also modeling the association between the two outcomes using the correlation coefficient of the bivariate latent variable, conditional on random effects. The model provides parameters that are subject specific but also have a population-averaged interpretation when scaled appropriately.

EXERCISES

10.1. Use a random effects model to estimate the effect in the subject-specific model that Section 8.2.5 considered for opinions about government performance in providing health care and in protecting the environment. Compare the results to the ones shown there.

10.2. Refer to Exercise 8.5 with Table 8.11 and Example 10.5.2 on opinions about government spending.

    (a) Formulate a cumulative logit random intercept model. Show that the ML estimates are as Section 10.5.2 shows. Interpret.

    (b) Fit the corresponding marginal cumulative logit model, using ML or GEE. Report estimates, explaining why they tend to be smaller than the estimates for the random effects model.

    (c) For an adjacent-categories logit random intercept model, when $\beta_4 = 0$, the ML estimates are $\hat{\beta}_1 = -1.92$ (SE $= 0.11$), $\hat{\beta}_2 = -0.37$ (SE $= 0.10$), and $\hat{\beta}_3 = -0.06$ (SE $= 0.11$), with $\hat{\sigma}_u = 0.61$ for the random effects. Interpret, and explain why the estimates tend to be smaller than for the model in part (a).

10.3. Explain how you could use a random effects model to analyze effects of explanatory variables on severity of retinopathy for the study described in Section 9.1.3, by using different variance components for the random effects for females and for males to handle the different association between the two responses that seems to occur for the two genders.

10.4. Refer to Table 9.5 from the insomnia study.

    (a) Analyze these data with a random effects model. Report and interpret the effects.

    (b) Fit the simpler model forcing $\sigma_u = 0$. Compare models, using either a likelihood-ratio test or AIC. What do you conclude?

    (c) Analyze the data with a marginal model. Compare results and interpretations to those in part (a).

CHAPTER 11

Bayesian Inference for Ordinal Response Data

This book has taken the traditional, so-called *frequentist*, approach to statistical inference. In this approach, probability statements apply to possible values for the data, given the parameter values, rather than to possible values for the parameters, given the data. Recent years have seen increasing popularity of an alternative approach, the *Bayesian* approach, which has probability distributions for parameters as well as for data and which assumes a *prior distribution* for the parameters which may reflect prior beliefs or information about the parameter values. This information is combined with the information that the data provide through the likelihood function to generate a *posterior distribution* for the parameters.

With the Bayesian approach, the data analyst makes probability statements about parameters, given the observed data, instead of probability statements about hypothetical data given particular parameter values. This results in inferential statements such as "Having observed these data, the probability is 0.03 that the population cumulative log odds ratio is negative" instead of "If the population cumulative log odds ratio were 0, then in a large number of randomized studies of this type, the proportion of times that we would expect to observe results more positive than the ones we observed is 0.03." The Bayesian approach seems more natural to many researchers, but it requires adding further structure to the model by the choice of the prior distribution.

In this chapter we apply the Bayesian paradigm to analyses of ordered categorical response data. In Section 11.2 we focus on Bayesian estimation of cell probabilities for a multinomial distribution with ordered categories or for a contingency table in which at least one variable is ordinal. In Section 11.3 we focus on Bayesian regression modeling of ordinal response variables, with emphasis on cumulative link models. In Section 11.4 we focus on association modeling and in Section 11.5 cite extensions to multivariate response models. Some general comparisons of Bayesian and frequentist approaches to analyzing ordinal data are

made in Section 11.6. First, though, we summarize the basic ideas that underlie the Bayesian approach.

11.1   BAYESIAN APPROACH TO STATISTICAL INFERENCE

As mentioned above, the Bayesian approach treats a parameter as a random variable having a probability distribution rather than as a fixed (but unknown) number. Statistical inference relies on the posterior distribution of a parameter, given the data, rather than the likelihood function alone.

11.1.1   Role of the Prior Distribution

To obtain the posterior distribution, we must first choose a prior distribution. The approach in doing so may be subjective or objective. In the subjective approach, the prior distribution reflects the researcher's prior information (before seeing the data) about the value of the parameter. The prior information might be based on eliciting beliefs about effects from the scientist who planned the study, those beliefs perhaps being based on research results available in the scientific literature. In the objective approach, the prior distribution might be chosen so that it has little influence on the posterior distribution, and resulting inferences are substantively the same as those obtained with a frequentist analysis. This might be done if the researcher prefers the Bayesian approach mainly because of the interpretive aspect of treating the parameter as random rather than fixed.

Introduction of the prior distribution to the statistical analysis is the key aspect of the Bayesian approach that is not a part of frequentist inference. Different choices for the prior distribution can result in quite different ultimate inferences, especially for small sample sizes, so the choice should be given careful thought. In this chapter we present the families of prior distributions commonly used for multinomial parameters and for parameters in ordinal models for multinomial responses.

The method of combining a prior distribution with the likelihood function to obtain a posterior distribution is called *Bayesian* because it is based on applying *Bayes' theorem*. By that result, the posterior probability density function h of a parameter θ, given the data $\mathbf{y}$, relates to the probability mass function f of $\mathbf{y}$, given θ, and the prior density function g of θ, by

$$h(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)g(\theta)}{f(\mathbf{y})}.$$

The denominator $f(\mathbf{y})$ on the right-hand side is the marginal probability mass function of the data. This is a constant with respect to θ, so is irrelevant for inference about θ. Computational routines must determine it so that the posterior density function $h(\theta \mid \mathbf{y})$ integrates to 1 with respect to θ, yielding a legitimate probability distribution for θ. When we plug in the observed data, $f(\mathbf{y} \mid \theta)$ is the likelihood function when viewed as a function of θ. So, in summary, the prior

density function multiplied by the likelihood function determines the posterior distribution.

Unlike the frequentist approach, the Bayesian approach does not differentiate between large- and small-sample analyses. Inference relies on the same posterior distribution regardless of the sample size n. Under standard regularity conditions such as θ not being on the boundary of the parameter space, as n grows the posterior distribution approximates more closely a normal distribution. Similarly, as n grows, a Bayesian estimator of θ (usually, the mean of the posterior distribution) behaves in a manner similar to the maximum likelihood estimator in terms of properties such as asymptotic normality, consistency, and efficiency. For example, Freedman (1963) showed consistency of Bayesian estimators under general conditions for sampling from discrete distributions such as the multinomial. He also showed asymptotic normality of the posterior distribution, assuming a smoothness condition for the prior distribution.

11.1.2   Simulating the Posterior Distribution

Except in specialized cases such as those presented in Section 11.2, there is no closed-form expression for the posterior distribution. Simulation methods can then approximate the posterior distribution. The primary method for doing this is Markov chain Monte Carlo (MCMC), with particular versions of it being the Metropolis–Hastings algorithm and its special case of Gibbs sampling. It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, a stochastic process of Markov chain form is designed so that its long-run stationary distribution is the posterior distribution. One or more such Markov chains provide a very large number of simulated values from the posterior distribution, and the distribution of the simulated values approximates the posterior distribution.

The process begins by the data analyst selecting initial estimates for the parameters and using a "burn-in period" for the Markov chain until its probability distribution is close to the stationary distribution. After the burn-in period, the simulated values are treated as providing information about the posterior distribution. The successive observations from the Markov chain are correlated, but observations that are a sufficient number of lags apart have little correlation. So "thinning" the process by taking lagged values (such as every fourth value) provides essentially uncorrelated observations from the posterior distribution. Enough observations are taken after the burn-in period so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of interest, such as the posterior mean and certain quantiles.

Various graphical and numerical summaries provide information about when the stationary condition has been reached and the Monte Carlo error is small. For a given parameter, plotting the simulated values against the iteration number is helpful for showing when burn-in is sufficient. After that, plotting the mean of the simulated values since the burn-in period against the iteration number helps to show when the approximation for the posterior mean has stabilized. A plot of the autocorrelations of lagged simulated values indicates how much the values must

be thinned to achieve approximately uncorrelated simulations from the posterior distribution. An alternative approach does not use thinning but accounts for the correlations.

Software is now widely available that can perform these computations for the basic analyses presented in this chapter. Textbooks specializing in computational aspects of the Bayesian approach, such as Ntzoufras (2009), provide details.

11.1.3 Inference Using the Posterior Distribution

Having found the posterior distribution, Bayesian methods of inference are available that parallel those for frequentist inference. For example, to summarize the information about a parameter β, we can report the mean and standard deviation of the posterior distribution of β. We can also construct an interval that contains most of that distribution. Analogous to the frequentist confidence interval, it is often referred to as a *credible interval* or *Bayesian confidence interval*.

A common approach for constructing a credible interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% credible interval for β is the region between the 2.5 and 97.5 percentiles of the posterior distribution for β. An alternative approach constructs a highest posterior density (HPD) region. This is the region of β values such that the posterior density function everywhere over that region is higher than the posterior density function everywhere outside that region, and the posterior probability over that region equals 0.95. Usually the HPD region is an interval, the narrowest one possible with the chosen probability.

A caution: For a parameter β in the linear predictor of a model, it is not sensible to construct a HPD interval for a nonlinear function of β. For a cumulative logit model, for instance, an HPD interval makes sense for β but not for $\exp(\beta)$. To illustrate, suppose that β is the coefficient of a binary indicator for two groups. Then a HPD interval for the cumulative odds ratio $\exp(\beta)$ does not consist of the reciprocals of values of the HPD interval for $\exp(-\beta)$, which is the cumulative odds ratio for the reverse assignment of values for the indicator variable. This is because the HPD region for a random variable $1/Z$ *is not* the same as the set of reciprocals of values in the HPD region for the random variable Z. However, the HPD region for $-Z$ *is* the same as the set of negatives of values in the HPD region for Z, so an HPD interval for β (a log cumulative odds ratio) is valid.

In lieu of P-values, with the Bayesian approach, posterior tail probabilities are useful. For example, for an effect parameter β, the information about the effect direction is contained in the posterior probabilities $P(\beta > 0 \mid \mathbf{y})$ and $P(\beta < 0 \mid \mathbf{y})$. With a flat prior distribution, $P(\beta > 0 \mid \mathbf{y})$ corresponds to the frequentist P-value for the one-sided test with $H_a: \beta < 0$.

Instead of a formal hypothesis test comparing models or possible values for a parameter, it is often useful to construct a *Bayes factor*. Consider data $\mathbf{y}$ and two models M_1 and M_2, which could also be two possible ranges of parameter values for a given model, two hypotheses, or even two nonnested distinct model types. The Bayes factor is the ratio of the marginal probabilities, $P(\mathbf{y} \mid M_2)/P(\mathbf{y} \mid M_1)$. This ratio gives the amount of evidence favoring one model relative to the other.

Many other analyses, including ways of checking models and an analog of AIC (called BIC), are available but are beyond the scope of this brief chapter treatment. For more in-depth presentations of Bayesian methods, particularly as they apply to ordered categorical data, see Johnson and Albert (1999), O'Hagan and Forster (2004, Sec. 12.52), and Congdon (2005, Chap. 7).

11.2 ESTIMATING MULTINOMIAL PARAMETERS

We first present Bayesian methods for estimating parameters of a multinomial distribution. The distribution could refer to counts for categories of an ordinal variable or to cell counts in a contingency table in which at least one variable is ordinal.

11.2.1 Estimation Using Dirichlet Prior Distributions

We begin with the multinomial with $c = 2$ categories, that is, the binomial. The simplest Bayesian inference for a binomial parameter π uses a member of the *beta distribution* as the prior distribution. The beta distribution is called the *conjugate prior distribution* for inference about a binomial parameter. This means that it is the family of probability distributions such that when combined with the likelihood function, the posterior distribution falls in the same family. When we combine a beta prior distribution with a binomial likelihood function, the posterior distribution also is a beta distribution. The beta probability density function for π is proportional to $\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$ for two parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. The family of beta probability density functions has a wide variety of shapes over the interval (0, 1), including uniform (when $\alpha_1 = \alpha_2 = 1$), unimodal symmetric ($\alpha_1 = \alpha_2 > 1$), unimodal skewed left ($\alpha_1 > \alpha_2 > 1$), unimodal skewed right ($\alpha_2 > \alpha_1 > 1$), and bimodal U-shaped ($\alpha_1 < 1, \alpha_2 < 1$).

For $c > 2$ categories, the beta distribution generalizes to the *Dirichlet distribution*. Its probability density function is positive over the simplex of nonnegative values $\pi = (\pi_1, \ldots, \pi_c)$ that sum to 1. Expressed in terms of gamma functions and parameters $\{\alpha_i > 0\}$, it is

$$g(\pi) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{c} \pi_i^{\alpha_i-1} \quad \text{for } 0 < \pi_i < 1 \text{ all } i, \quad \sum_i \pi_i = 1.$$

The case $\{\alpha_i = 1\}$ is the uniform distribution over the simplex of possible probability values. The case $\{\alpha_i = \frac{1}{2}\}$ is the *Jeffreys prior* for the multinomial distribution, which is the prior distribution that is invariant to the scale of measurement for the parameter. Let $K = \sum_i \alpha_i$. The Dirichlet distribution has $E(\pi_i) = \alpha_i/K$ and $\text{Var}(\pi_i) = \alpha_i(K - \alpha_i)/[K^2(K + 1)]$. For particular relative sizes of $\{\alpha_i\}$, such as identical values, the distribution is more tightly concentrated around the means as K increases.

For cell counts $\mathbf{n} = (n_1, \ldots, n_c)$ from $n = \sum_i n_i$ independent observations with cell probabilities $\boldsymbol{\pi}$, the multinomial probability mass function is

$$f(\mathbf{n} \mid \boldsymbol{\pi}) = \frac{n!}{n_1! \cdots n_c!} \prod_{i=1}^{c} \pi_i^{n_i}.$$

Given $\mathbf{n}$, multiplying this multinomial likelihood function by the Dirichlet prior density function $g(\boldsymbol{\pi})$ yields a posterior density function $h(\boldsymbol{\pi} \mid \mathbf{n})$ for $\boldsymbol{\pi}$ that is also a Dirichlet distribution, but with the parameters $\{\alpha_i\}$ replaced by $\{\alpha_i' = n_i + \alpha_i\}$. The mean of the posterior distribution of π_i is

$$E(\pi_i \mid n_1, \ldots, n_c) = \frac{n_i + \alpha_i}{n + K}.$$

Let $\gamma_i = E(\pi_i) = \alpha_i / K$. This Bayesian estimator equals the weighted average

$$\frac{n}{n + K} p_i + \frac{K}{n + K} \gamma_i \tag{11.1}$$

of the sample proportion $p_i = n_i / n$ and the mean γ_i of the prior distribution for π_i. This posterior mean takes the form of a sample proportion when the prior information corresponds to K additional observations of which α_i were outcomes of type i.

Articles by Lindley (1964) and Good (1965) were influential early works that estimated multinomial parameters using a Bayesian approach with a Dirichlet prior distribution. Good's book and earlier related articles were motivated by work he conducted with Alan Turing at Bletchley Park, England, during World War II in breaking the German code used for wartime communications. For Dirichlet prior distributions with identical $\{\alpha_i\}$, Good referred to K as a *flattening constant*, because the Bayes estimate shrinks each sample proportion toward the equiprobability value $\gamma_i = 1/c$. Greater shrinkage occurs as K increases, for fixed n.

In summary, the Bayesian estimators shrink the sample proportions, which are the unrestricted ML estimates, toward their prior means. The Bayesian estimators combine good characteristics of sample proportions and of frequentist model-based estimators, such as the ML estimator $1/c$ for the simple equiprobability model. Like sample proportions and unlike model-based estimators, the Bayesian estimators are consistent even when a particular model (such as the equiprobability model) does not hold. The weight given the sample proportion then increases to 1.0 as the sample size increases. Like model-based estimators and unlike sample proportions, the Bayesian estimators smooth the data. For example, cells with no observations have Bayesian probability estimates that are strictly positive.

Unlike the sample proportions, Bayesian estimators of multinomial parameters are slightly biased for finite n. Usually, though, they have smaller total mean-squared error (MSE) than the sample proportions. They are not uniformly better for all possible parameter values, however. For example, if a particular $\pi_i = 0$, then $p_i = 0$ with probability 1, so the sample proportion is then better than any

other estimator. We do not expect $\pi_i = 0$ in practice, but this limiting behavior explains why the ML estimator can have smaller MSE than the Bayes estimator when π_i is very near 0.

11.2.2 Example: Government Spending on the Arts

The 2006 General Social Survey asked respondents how much government should spend on culture and the arts, with categories (much more, more, the same, less, much less). Table 11.1 shows the results for 18 to 21-year-old females. None of the 23 females in this sample said "much less," so the ML estimate for that category is the sample proportion value of 0.0. It is implausible that the corresponding population proportion is 0, and it is sensible to smooth these estimates.

Table 11.1 also shows Bayes' estimates of the population proportions based on a Dirichlet prior with $\{\alpha_i = K/5\}$ for various values of the flattening constant K. The prior with $K = 1$ provides only slight smoothing, the empty cell still having an estimated probability below 0.01. Greater smoothing occurs with the uniform prior, for which each $\alpha_i = 1$ and $K = 5$. The estimates then correspond to sample proportions after adding an observation to each cell, increasing the effective sample size from 23 to 28. Supersmoothing occurs with the choice $K = 20$, by which the prior information receives nearly as much weight as the data.

11.2.3 Smoothing Contingency Tables

This Bayesian analysis extends to estimating multinomial probabilities that result from cross-classifying categorical variables in a contingency table. Denote the cell counts by $\mathbf{n} = \{n_i\}$ and the cell probabilities by $\boldsymbol{\pi} = \{\pi_i\}$, where these can refer to a table of any dimension. This section does not distinguish between response and explanatory variables (Section 11.3 does this) but focuses instead on smoothing the data to improve the estimation of $\boldsymbol{\pi}$.

With a Dirichlet prior distribution for $\boldsymbol{\pi}$, the posterior mean estimate for a particular cell probability again has the form (11.1). That is, the Bayesian estimator shrinks the sample proportions $\mathbf{p} = \{p_i\}$ toward the prior means.

TABLE 11.1. Opinions About Government Spending on the Arts

| Estimates | Government Spending | | | | |
	Much more	More	Same	Less	Much less
Sample counts	1	7	12	3	0
Sample proportions	0.043	0.304	0.522	0.130	0.000
Bayes, $K = 1$	0.050	0.300	0.508	0.133	0.008
Bayes, $K = 5$ (uniform)	0.071	0.286	0.464	0.143	0.036
Bayes, $K = 20$	0.116	0.256	0.372	0.163	0.093

Source: 2006 General Social Survey.

Sometimes it can be useful to shrink the sample proportions toward a particular model, such as independence between two variables. This can be effective even when we do not think that model truly holds but we believe that the parsimonious aspects of the model are beneficial for smoothing the data. To do this, we could choose values for the prior means that satisfy the model.

These two versions of Bayesian estimation do not account for the ordinality of a categorical variable. Without further structure, the analyses do not differentiate between ordinal variables and nominal variables in the contingency table. To illustrate, consider the smoothing above of the multinomial counts in Table 11.1 on opinions about government spending on the arts. For a given K, the same estimates occur if the categories are permuted in any way.

One way to extend this approach so that it recognizes the ordering of categories is to let the prior means satisfy an ordinal model. For example, suppose that the contingency table is a two-way cross-classification of two ordinal variables. Then we could let $\{\gamma_{ij} = E(\pi_{ij})\}$ satisfy an ordinal association model such as a linear-by-linear association. To do this we could construct prior best guesses for cell probabilities that have a monotone trend, smooth them by fitting the ordinal model to them, and use those fitted values as the prior means.

11.2.4 Empirical Bayesian Estimates of Multinomial Parameters

Rather than requiring the data analyst to select parameter values for the prior distribution, another approach uses the data to determine those values. This is called the *empirical Bayesian* approach. With a common version of this approach, the estimated prior distribution is the one that maximizes the marginal probability $f(\mathbf{y})$ of the observed data, integrating out with respect to the prior distribution. Good (1965) used this approach to estimate the parameter value for a symmetric Dirichlet prior distribution for multinomial parameters.

Fienberg and Holland (1973) proposed alternative estimates of multinomial probabilities $\{\pi_i\}$ that use data-dependent priors. For a particular choice of Dirichlet means $\{\gamma_i\}$ for the Bayesian estimator

$$\frac{n}{n+K} p_i + \frac{K}{n+K} \gamma_i,$$

they showed that the minimum total mean squared error occurs when

$$K = \frac{1 - \sum_i \pi_i^2}{\sum_i (\gamma_i - \pi_i)^2}.$$

The optimal $K = K(\gamma, \pi)$ depends on π, and they used the estimate $\hat{K} = K(\gamma, \mathbf{p})$ from substituting the sample proportion $\mathbf{p}$ for π. As $\mathbf{p}$ falls closer to the prior guess γ, $\hat{K}$ increases and the prior guess receives more weight in the posterior estimate. They selected $\{\gamma_i\}$ based on the fit of a simple model. For elaboration and extensions, see Bishop et al. (1975, Chap. 12).

We illustrate for the opinions about government spending on the arts in Table 11.1. Section 11.2.2 estimated the cell probabilities using Dirichlet priors with $\gamma_i = \frac{1}{5}$ for the five categories. With the optimal K estimated by substituting the sample proportions, for this prior with identical prior parameters we obtain $\hat{K} = (1 - 0.384)/0.184 = 3.35$, not much different from the uniform prior ($K = 5$). The corresponding empirical Bayesian estimates of the five probabilities are (0.063, 0.291, 0.481, 0.139, 0.025), compared to the sample proportions of (0.043, 0.304, 0.522, 0.130, 0.000).

For two-way tables with sample cell proportions $\{p_{ij}\}$, Fienberg and Holland estimated the prior means by the fit of the independence model, $\{\gamma_{ij} = p_{i+}p_{+j}\}$. When the categories in at least one dimension in a contingency table are ordered, we would usually improve on this by using the fit of an ordinal model to specify $\{\gamma_{ij}\}$. Agresti and Chuang (1989) used this approach by smoothing toward the linear-by-linear association model (6.2). They also considered an approach to smoothing that uses a Bayesian analysis for the saturated loglinear model but has linear-by-linear structure for the prior means of the association parameters.

11.2.5 Example: Government Spending on Health, by Gender

In Section 11.2.2 we analyzed GSS responses from 2006 about government spending on the arts. For the same year, Table 11.2 shows opinions about government spending on health, by gender, for subjects of age 18 to 25. The table also shows the sample conditional distributions on the opinion response variable. The two groups are stochastically ordered, females showing a slight tendency to fall more toward the "spend much more" end of the scale. With such a small sample, sample percentage estimates are unappealing. For example, it is implausible that the population percentage of females who favor spending less or much less on health is 0.0. At the same time, there is not enough data to put much faith in any particular model for the association. As a compromise, it seems sensible to smooth the sample percentage estimates by shrinking them toward the fit of an ordinal model.

TABLE 11.2. Opinions About Government Spending on Health, by Gender, with Sample Conditional Distribution on Spending and Bayes Estimates Based on Shrinking Toward Ordinal Model

		Spending on Health					
		Much More	More	Same	Less	Much Less	Total
Female	Count	30	26	8	0	0	64
	Percentage	46.9%	40.6%	12.5%	0.0%	0.0%	100%
	Bayes estimate	48.1%	41.2%	8.4%	1.7%	0.7%	100%
Male	Count	25	28	3	4	2	62
	Percentage	40.3%	45.2%	4.8%	6.5%	3.2%	100%
	Bayes estimate	39.0%	44.6%	9.1%	4.7%	2.5%	100%

Source: 2006 General Social Survey.

Despite the stochastic ordering, ordinal logit models of proportional odds form that have an indicator variable for gender show some lack of fit. For example, the adjacent-categories logit model, which is equivalent to the loglinear model of uniform association for the local odds ratios (i.e., linear-by-linear association with equally spaced scores) has deviance 9.2 (df = 3). Although we might prefer not to use the fit of this model to estimate the cell probabilities, we could use the model as a mechanism for smoothing the sample percentages. This takes into account the ordering of the opinion categories, unlike Bayesian smoothing with ordinary Dirichlet priors.

For the Fienberg–Holland data-based prior approach with the sample of size $n = 126$, we obtain $\hat{K} = 420.45$. The estimated percentages are then weighted averages with weight $n/(n + \hat{K}) = 0.23$ given to the sample percentage and weight $\hat{K}/(n + \hat{K}) = 0.77$ given to the model fit. So there is strong shrinkage of the sample percentages toward the model fit. Table 11.2 also shows the corresponding Bayesian estimates of the conditional distributions on the opinion response variable. The sample percentages of 0.0 for females on the "spend less" and "spend much less" categories are replaced by the Bayesian estimates 1.7% and 0.7%.

11.2.6   Hierarchical Bayesian Estimates of Multinomial Parameters

Good (1965) presented yet another alternative to specifying the parameters of a Dirichlet prior distribution. This approach is a *hierarchical* one that specifies distributions also for the Dirichlet parameters $\{\alpha_i\}$. That is, Good treated $\{\alpha_i\}$ as unknown and specified a second-stage prior distribution for them. For example, in the symmetric case with common value $\{\alpha_i = \alpha\}$, in one example (p. 38) he assumed that $\log \alpha$ has a symmetrical distribution about 0 such that $P(\epsilon < \alpha < 1/\epsilon) = 1 - \epsilon$ for all $0 < \epsilon < 1$. This corresponds to a probability density function for α of

$$g(\alpha) = \frac{1}{2} \ \text{ for } \ 0 < \alpha < 1 \ \text{ and } \ g(\alpha) = \frac{1}{2\alpha^2} \ \text{ for } \ \alpha \geq 1.$$

More generally, the second-stage prior distribution can have its own parameters, called *hyperparameters*.

Epstein and Fienberg (1992) suggested an alternative specification of two-stage prior distributions for the cell probabilities of a contingency table. The first stage again places a Dirichlet(K, γ) prior on π. The second stage uses a loglinear model for the prior means $\{\gamma_i\}$, assuming a multivariate normal prior distribution on the loglinear terms. Applying the loglinear parameterization to the prior means $\{\gamma_i\}$ rather than directly to the cell probabilities $\{\pi_i\}$ permits the analysis to reflect uncertainty about the loglinear structure for $\{\pi_i\}$.

The hierarchical approach provides greater generality at the expense of not having the simple conjugate Dirichlet form for the posterior distribution. Computations are therefore more complex. Compared to the hierarchical approach, a disadvantage of the empirical Bayesian approach is that it does not account for the additional source of variability due to substituting estimates for prior parameters. In recent years, use of the hierarchical approach has increased, as it provides a

direct mechanism for representing uncertainty about the parameters of the prior distribution.

11.2.7   Estimation Using Logistic-Normal Prior Distributions

The Dirichlet distribution is sometimes not sufficiently flexible. Although we can specify the means $\{E(\pi_i)\}$ in the prior distribution through the choice of $\{\gamma_i\}$ and the variances through the choice of K, there is no freedom to alter the correlations among the parameters in that prior. In addition, we've seen that the Dirichlet prior distribution does not take into account the ordering of the categories for ordinal responses unless we put some ordinal structure on the prior means. Also, it does not naturally extend to parameters in models for the multinomial parameters, such as cumulative link models and hierarchical models.

An alternative to a Dirichlet prior distribution is a prior distribution induced by a multivariate normal distribution for multinomial logits such as the baseline-category logits. This family of distributions also lends itself more naturally than the Dirichlet to extensions such as hierarchical modeling. The corresponding prior distribution for the multinomial parameters themselves is the multivariate *logistic-normal distribution*. Specifically, if $\mathbf{X} = (X_1, \ldots, X_c)$ has a multivariate normal distribution, then $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$ with

$$\pi_i = \frac{\exp(X_i)}{\sum_{j=1}^{c} \exp(X_j)}$$

has a multivariate logistic-normal distribution. Specifying a mean vector and covariance matrix for $\mathbf{X}$ induces a particular prior distribution for $\boldsymbol{\pi}$. A more general version uses a hierarchical approach that also specifies prior distributions for the parameters of the normal distribution, as illustrated in the next example.

The logistic-normal prior distribution provides more flexibility than the Dirichlet prior distribution. For instance, with ordered categories we often expect probabilities in adjacent categories to be similar. One way to represent this uses an autoregressive form for the normal correlation matrix. The correlation between X_a and X_b for categories a and b then has the form ρ^{a-b} for $|\rho| < 1$, for which X_a and X_b are more strongly correlated for categories that are closer together. Using the logistic-normal prior distribution also connects inference here with Bayesian inference for parameters in logistic regression models using normal prior distributions for the model parameters.

11.2.8   Example: Smoothing a Histogram for Apple Diameter

Leonard (1973) used a logistic-normal prior distribution for the purpose of smoothing a histogram, treating the counts in the various class intervals as having a multinomial distribution. When the intervals have equal width, probabilities in adjacent intervals are often similar. An ordinary histogram uses the sample proportion for category j to estimate the probability for that category. By contrast,

the Bayesian approach uses the information from all the categories to estimate the probability for any one category.

Leonard applied this method of smoothing a histogram to data on the maximum diameters of apples of a certain species. The diameters were recorded in 10 intervals, (2.0, 2.1], (2.1, 2.2], ... , (2.9, 3.0]. Table 11.3 shows the counts and sample proportions for these intervals for a sample of 152 apples. The ordinary histogram using the sample proportions has modes for the intervals 1, 4, 6, 8. It is rough rather than smooth. Leonard found Bayesian estimates of the probabilities in these ten intervals using a hierarchical approach with a two-stage prior distribution. The first stage is a logistic-normal prior distribution for the probabilities. For the covariance structure of the multivariate normal, he let categories a and b have correlation of the form ρ^{a-b}, reflecting categories that are closer together tending to have more similar probabilities. Leonard assumed no prior information about which categories would have higher probabilities. Thus, for the multivariate normal vector for which the logits generate the probabilities, he took the elements of the mean vector to be identical and elements σ^2 of the variance vector to be identical. This corresponds to values of 1/10 for prior means for each probability.

Specifying ρ and σ^2 determines the logistic-normal prior distribution, and Bayesian methods generate the posterior distribution. However, Leonard suggested adding a second stage to the prior distribution, giving prior distributions also to ρ and $\tau = 1/\sigma^2(1 - \rho^2)$. He let $\nu\lambda\tau$ have a chi-squared distribution with df $= \nu$, for specified values of ν and λ, where λ is a prior estimate of τ^{-1} and ν measures the sureness of this estimate. For technical convenience, he let ρ have a normal distribution, with specified mean and variance, truncated over the interval $[-1, 1]$. He found that for hyperparameter values corresponding to very flat distributions for ρ and σ^2, results were not sensitive to the choice of those values.

Table 11.3 also shows Bayesian estimates of the multinomial probabilities based on Leonard's analysis. Unlike the histogram of sample proportions, the smoothed histogram has just one mode and has a more appealing shape. The posterior mode for ρ equals 0.77, reflecting relatively strong association for adjacent categories. However, Leonard also noted that the Bayes estimates in categories 1, 2, 9, and 10 being greater than the sample proportions was partly due to having equal prior means for those categories. If it were considered more reasonable to have smaller values in the tails, that prior structure may not be sensible.

TABLE 11.3. Frequency Distribution for Ten Intervals for a Histogram for Apple Diameter

Interval Summary	Apple Diameter Interval									
	1	2	3	4	5	6	7	8	9	10
Sample count	5	4	10	18	15	30	25	27	12	6
Sample proportion	0.033	0.026	0.066	0.118	0.099	0.197	0.164	0.178	0.079	0.039
Bayes estimate	0.042	0.045	0.067	0.102	0.111	0.172	0.161	0.153	0.086	0.060

Source: Based on analysis presented in Leonard (1973).

With Leonard's approach, letting $\rho = 0$ provides an exchangeable structure among the categories that ignores the ordering and is analogous to the structure generated by a symmetric Dirichlet prior distribution. In that case, the Bayesian estimate of the logit for category j is approximately equal to a weighted average of the sample logit and the prior mean for the logit, with weight proportional to σ^{-2} for the prior mean. So, with the logistic-normal prior distribution, the inverse variance of each normal component serves as the analog of Good's flattening coefficient, with smaller σ^2 giving more weight to the prior mean and resulting in greater shrinkage.

11.3   BAYESIAN ORDINAL REGRESSION MODELING

Bayesian modeling of ordinal categorical response variables provides an alternative to the frequentist modeling of Chapters 3 to 5. Our main focus here is on cumulative link models, such as the cumulative logit and probit models.

11.3.1   Improper and Other Relatively Flat Prior Distributions

Models can have many parameters, and a researcher may have more prior information about some of them than others. It can be challenging to specify a sensible prior distribution, especially when the parameters relate to nonlinear transformations of probabilities such as a cumulative logit. One simplistic approach takes the prior distribution to be constant over the multidimensional space of all possible parameter values. Then the posterior distribution is a constant multiple of the likelihood function. That is, the posterior distribution is a scaling of the likelihood function so that it integrates out to 1. The mode of the posterior distribution is then the ML estimate.

When some parameters can take value over the entire real line, such as β effect parameters in cumulative link models, such a flat prior distribution is said to be *improper*. It is not a legitimate probability distribution, because it does not integrate out to 1 over the space of possible parameter values. A danger is that improper prior distributions also have improper posterior distributions for some models (Hobert and Casella 1996). A Markov chain Monte Carlo (MCMC) algorithm for approximating the posterior distribution may fail to recognize that the posterior distribution is improper. Thus, it is safer to use a proper but relatively diffuse prior if you prefer the prior distribution to be flat relative to the likelihood function.

One such diffuse prior for a given parameter is a normal distribution with a very large standard deviation. However, when there are many parameters, the posterior mode need not then necessarily be close to the ML estimate, and Markov chains may converge slowly (Natarajan and McCulloch 1998). The mean can be quite different from the mode when the sample size is small or the data are unevenly distributed among the categories, in which case the posterior distribution may be quite skewed rather than approximately normal. When you use the Bayesian approach, it is usually more sensible to construct prior distributions that represent careful

expression of prior beliefs about the parameter values. For example, instead of using a very large standard deviation for a normal prior distribution, use a mean and standard deviation such that the range within three standard deviations of the mean contains all values that are at all plausible for the parameter.

11.3.2 Proper Prior Specifications for Cumulative Link Models

For frequentist inference with an ordinal response variable, we have seen that many models are special cases of the cumulative link model (5.1), which with link function h is

$$h[P(Y_i \le j)] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}_i. \tag{11.2}$$

From Section 3.3.2, such ordinal models are implied by standard regression models for underlying latent variables, such as logistic for the logit link and normal for the probit link. The model with the negative sign attached to the predictor effects naturally results from assuming a standardized distribution for $(y^* - \mu)/\sigma$ when a latent variable y^* has mean $\mu = \boldsymbol{\beta}'\mathbf{x}$ and standard deviation σ at setting $\mathbf{x}$ for values of explanatory variables. We use this parameterization because the model-fitting described below in Section 11.3.3 refers to the latent variable model.

Prior distributions for parameters in cumulative link models should take into account the ordering constraint

$$-\infty < \alpha_1 < \alpha_2 < \cdots < \alpha_{c-1} < \infty$$

on the cutpoint parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{c-1})$. Posterior distributions for the parameters $\boldsymbol{\beta}$ of interest mainly depend on the choice of prior for $\boldsymbol{\beta}$, so the prior for $\boldsymbol{\alpha}$ is taken to be relatively diffuse. We shall assume that no parameters in the linear predictor $\alpha_j - \boldsymbol{\beta}'\mathbf{x}_i$ are redundant (e.g., $\boldsymbol{\beta}$ does not contain a separate intercept term), so it is not necessary to add further constraints. Considerable flexibility is provided by a multivariate normal prior density, plausible values for a parameter being dictated by the choice of prior mean and standard deviation. For example, for the cumulative probit model, Chipman and Hamada (1996) used a multivariate normal prior distribution for $\boldsymbol{\beta}$ and a truncated multivariate normal prior distribution for $\boldsymbol{\alpha}$ that respects the ordering of their values. They illustrated with two industrial data sets. If appropriate, you can include correlation in the prior distribution between different parameters.

In practice, meaningful specification of the parameters for such normal prior distributions for $\boldsymbol{\beta}$ is not obvious, because data analysts think more easily in terms of plausible values for probabilities rather than for model parameters that pertain to a nonlinear function of the cumulative probabilities such as cumulative log odds ratios. Alternatively, you can construct a prior distribution on the cumulative probability scale rather than a link function scale such as the logit. The chosen prior distribution then induces a corresponding prior distribution for the model parameters. Bedrick et al. (1996) proposed this approach in the binary case, and

Johnson and Albert (1999) extended it to ordinal responses. This approach requires prior distributions for at least as many cumulative probability values as there are parameters in the model. For example, if a cumulative logit model of proportional odds form has three explanatory variables for a three-category response variable, and hence five parameters, we need at least five such prior distributions.

Suppose we choose M settings of explanatory variable values for placing prior distributions on cumulative probabilities. At setting s, denoted by $\mathbf{x}_{(s)}$, we select a particular cumulative probability and formulate a prior distribution (such as a beta distribution) for it. Denote that cumulative probability by $\gamma_{(s)}$. One way to indirectly determine the parameters for a beta distribution is to guess the value of the cumulative probability (call that guess g_s) and to select a number K_s of "prior observations" that the prior belief represents. Then, the beta prior density for that cumulative probability has parameters $K_s g_s$ and $K_s(1 - g_s)$, corresponding to mean g_s. For example, we might predict that the first cumulative probability at a particular setting of the explanatory variable takes value 0.3 and that our prior information is relatively vague, say being comparable to two prior observations. We do this at all M settings.

For simplicity, these M prior distributions for the cumulative probabilities are assumed to be independent. Then the joint prior density function in terms of these M cumulative probabilities is

$$g(\gamma_1, \ldots, \gamma_M) \propto \prod_s \gamma_{(s)}^{K_s g_s - 1} (1 - \gamma_{(s)})^{K_s(1 - g_s) - 1}.$$

Suppose that the link function for the model corresponds to the inverse of the cumulative distribution function F, such as standard normal for the probit link and standard logistic for the logit link. Let f denote the corresponding probability density function. Then, in terms of the model parameters this prior density function corresponds to

$$g(\boldsymbol{\beta}, \boldsymbol{\alpha}) \propto \prod_s \{F(\alpha_{(s)} - \boldsymbol{\beta}'\mathbf{x}_{(s)})^{K_s g_s - 1}[1 - F(\alpha_{(s)} - \boldsymbol{\beta}'\mathbf{x}_{(s)})]^{K_s(1 - g_s) - 1}$$
$$\times f(\alpha_{(s)} - \boldsymbol{\beta}'\mathbf{x}_{(s)})\},$$

subject to the ordering constraint on $\boldsymbol{\alpha}$.

Other types of prior distributions have been proposed for ordinal regression models. For example, Gill and Casella (2009) proposed nonparametric prior distributions based on a Dirichlet process.

11.3.3 Bayesian Fitting of Cumulative Link Models

We now summarize the basic ideas of Bayesian model fitting of the cumulative link model, glossing over the technical details. Albert and Chib (1993) presented a Bayesian analysis for binary and multinomial models, implemented with Gibbs sampling, that has been influential. For cumulative probit models, it utilizes the

latent variable model shown in Section 3.3.2. This model is simpler to handle than the cumulative logit model, because results apply from Bayesian inference for ordinary normal linear regression models. They assumed a multivariate normal prior distribution for the regression parameters and independent normal latent variables. Then, the posterior distribution of the regression parameters, conditional on the observed data and the latent variables, is multivariate normal. Implementation of MCMC methods is relatively simple because the Monte Carlo sampling is from a normal distribution. The approach described here partly borrows from that article and from Johnson and Albert (1999, Chap. 6.2).

For the chosen prior specification, the prior density function is multiplied by the multinomial likelihood function for the model to obtain the posterior density function, apart from the proportionality constant needed so the posterior integrates to 1. Denote observation i by y_i. If $y_i = j$, its contribution to the likelihood function is

$$F(\alpha_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\alpha_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i).$$

This is simply the multinomial mass function for that single observation evaluated after observing the outcome, so it is the probability of category j.

Equivalently, as in Albert and Chib (1993), the likelihood function can be constructed in terms of the model for the underlying latent observation, y_i^*. If $y_i = j$, then y_i^* fell between α_{j-1} and α_j. So, for the latent observation with standardized density f, the contribution to the likelihood function (also regarding the latent observation as if it were an unknown parameter) is

$$f(y_i^* - \boldsymbol{\beta}'\mathbf{x}_i)I(\alpha_{j-1} \leq y_i^* \leq \alpha_j),$$

where I is the indicator function. Now, for the n independent observations, the likelihood function is proportional to

$$\prod_i f(y_i^* - \boldsymbol{\beta}'\mathbf{x}_i)I(\alpha_{y_i-1} \leq y_i^* \leq \alpha_{y_i}).$$

For uniform (improper) prior distributions, this is also the form of the posterior distribution, over the constrained space for $\boldsymbol{\alpha}$.

Using the ML estimates as initial values, Albert and Chib (1993) found the posterior distribution using a "one-run" hybrid MCMC algorithm with thinning that recognizes the ordering constraint on the $\{\alpha_j\}$ and uses highly diffuse prior distributions. Their Gibbs sampling scheme successively samples from the density of (1) $\mathbf{y}^*$ given $\mathbf{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$, (2) $\boldsymbol{\beta}$ given $\mathbf{y}$, $\mathbf{y}^*$, and $\boldsymbol{\alpha}$, and (3) $\boldsymbol{\alpha}$ given $\mathbf{y}$, $\mathbf{y}^*$, and $\boldsymbol{\beta}$. The model fitting results in posterior means for the Bayes estimates of any particular parameter. Posterior standard deviations describe the precisions of those estimates.

Albert and Chib also used a link function corresponding to the cdf of a t distribution, to investigate the sensitivity of fitted probabilities of response categories to the choice of link function. They noted that a t variate with df $= 8$ divided by 0.634 well approximates a standard logistic variate. The Cauchy link results with

df = 1 and the probit link results as df $\to \infty$. They considered the t link case through a latent variable model using a scale mixture of normal distributions. They also considered a hierarchical analysis that uses prior densities for the parameters of a normal prior distribution for β.

11.3.4 Example: Comparing Operations for Ulcer

We illustrate this Bayesian model-based analysis for the data in Table 11.4, from a study comparing two operations for duodenal ulcer. We use the cumulative logit model with proportional odds structure,

$$\text{logit}[P(Y_i \leq j)] = \alpha_j - \beta x_i,$$

where x_i is an indicator variable for the two operations.

For the Bayesian analysis, we used independent normal prior distributions. To reflect a lack of prior belief about the direction of the effect, we took the mean of the prior distribution of β to be 0.0. To reflect a lack of information about the size of the effect, we first took the prior distribution to be extremely diffuse, with a standard deviation of 1000. For the prior distributions of α_1 and α_2, we started with normal distributions having means of -1.0 for α_1 and $+1$ for α_2 and with standard deviations of 1000, but truncated the joint distribution so that $\alpha_1 < \alpha_2$. Instead of the usual $(0, 1)$ coding for the indicator variable x_i, we let it take value -0.5 if subject i has operation 1 and take value 0.5 if subject i has operation 2. The prior distribution is then symmetric in the sense that the cumulative logits in each row have the same prior variability as well as the same prior means, yet β still has the usual interpretation of a log cumulative odds ratio.

The analysis can be implemented with Bayesian software such as WinBUGS or SAS (PROC MCMC, as shown in Table A.8 in the Appendix), using a MCMC algorithm to approximate the posterior means, standard deviations, and quantiles. Table 11.5 shows posterior estimation results for β, based on an MCMC process using 1,000,000 iterations, with ML estimates as starting values and using the first 10,000 iterations for burn-in and with a thin value of 2. Chains were also run with other starting values, and gave similar results. With such a long process, the Monte

TABLE 11.4. Data from Clinical Trial Comparing Two Operations for Duodenal Ulcer

Treatment	Response		
	Death	Fair to Poor	Good to Excellent
Operation 1	7	17	76
Operation 2	1	10	89

Source: Adapted from M. Novick and J. Grizzle, *J. Amer. Statist. Assoc.* **60**: 91 (1965), with permission. Copyright © 1965 American Statistical Association. All rights reserved.

TABLE 11.5. Posterior Estimates of Cumulative Log Odds Ratio β for Analyses of Ulcer Data

β Prior Mean, Std. Dev.	Mean	Std. Dev.	95% Credible Interval	$P(\beta < 0)$
(0, 1)	0.836	0.361	(0.138, 1.555)	0.009
(0, 1000)	0.997	0.402	(0.229, 1.808)	0.005
ML^a	0.970	0.396	(0.216, 1.779)	0.006

[a]Results shown in ML row are the ML estimate, standard error, 95% profile likelihood confidence interval, and P-value for likelihood-ratio test with $H_a : \beta < 0$.

Carlo standard error for the approximation 0.997 to the Bayes estimate of β was negligible (about 0.001). The estimated odds of outcome in category j or below (for $j = 1, 2$) for operation 1 are $\exp(0.997) = 2.7$ times the estimated odds for operation 2. The 95% credible interval for β, based on the equal-tail method, is (0.23, 1.81). This provides the inference that $\beta > 0$. The estimated size of the effect is imprecise, because 4 of the 6 cell counts are quite small and the prior distributions were highly diffuse.

For comparison, Table 11.5 also shows results of the frequentist analysis. The cumulative logit model fits well, with Pearson statistic of $X^2 = 1.3$ on $df = 1$ for testing goodness of fit. The ML estimate of the common cumulative log odds ratio is $\hat{\beta} = 0.970$, with SE $= 0.395$ using the observed information matrix. The likelihood-ratio test of $H_0 : \beta = 0$ has chi-squared test statistic $= 6.42$. The P-value is 0.006 for the alternative $H_a : \beta < 0$. This test and the corresponding 95% confidence interval suggest that operation 1 is worse than operation 2. Inferences about β were substantively the same as with the Bayesian analysis. This is not surprising, because when the prior distribution is flat relative to the likelihood function, the posterior distribution itself is roughly proportional to the likelihood function.

For further comparison, we used a more informative prior distribution for β around the mean of 0.0. To reflect a belief that the size of the effect is not extremely strong, we took the prior standard deviation to be 1.0. Then nearly all the prior probability mass for the cumulative odds ratio $\exp(\beta)$ falls between $\exp(-3.0) = 0.05$ and $\exp(3.0) = 20$. For the prior distributions of α_1 and α_2, we started with normal distributions having means of -1.0 for α_1 and $+1$ for α_2 and with standard deviations of 2.0, but truncated the joint distribution so that $\alpha_1 < \alpha_2$. Such distributions can still accommodate a broad range of response prior probabilities. Results for estimating β were somewhat different than with the frequentist analysis or the Bayesian analysis with very flat prior distributions. The posterior mean for β is now 0.836 instead of 0.997. The prior standard deviation of 1.0 for β in this Bayesian analysis was not dramatically larger than the SE of 0.395 for $\hat{\beta}$ in the ML analysis, reflecting the small counts in four of the cells. As a consequence, the Bayesian analysis has non-trivial shrinkage of the ML estimate toward the prior mean of 0. The posterior mean is approximately the weighted average of the prior mean and the ML estimate,

$$\text{posterior mean} \approx (1 - w)(\text{prior mean}) + w(\text{ML estimate}),$$

with weights inversely proportional to the respective variances. The prior mean is 0 and the prior variance is 1.0, and the ML estimate of 0.970 has SE $= 0.395$, so the posterior mean is approximately

$$\frac{0.395^2}{1 + 0.395^2} (0.0) + \frac{1}{1 + 0.395^2} (0.970) = 0.84.$$

Corresponding to the frequentist P-value for H_a: $\beta > 0$, the Bayesian approach can report the posterior probability that $\beta < 0$. Table 11.5 also reports this probability for each choice of prior distributions. For the essentially flat prior distribution, this posterior tail probability of 0.005 is nearly identical to the P-value for the frequentist test of H_0: $\beta = 0$ against H_a: $\beta > 0$, thus giving very strong evidence that $\beta > 0$. With the more informative prior distribution centered at a lack of a treatment effect, this posterior probability provides slightly less evidence of a treatment effect.

11.3.5 Bayesian Fitting of Adjacent-Categories Logit Models

Adjacent-categories logit models for a multinomial response variable Y with categorical explanatory variables have corresponding Poisson loglinear models with equally-spaced scores for Y. For example, Section 6.2.3 noted that the uniform association model for a two-way contingency table is equivalent to an adjacent-categories logit model of proportional odds form. Assuming independent multinomial responses for Y corresponds to assuming independent Poisson observations for the cell counts in the contingency table and then conditioning on the row totals at the various combinations of levels of explanatory variables. Bayesian analyses for such adjacent-categories logit models can be conducted by performing Bayesian analyses for the corresponding loglinear models.

Using the loglinear connection has the advantage that Bayesian software that can handle Poisson responses can be used for Bayesian fitting of such models.[1] A simplification of using an adjacent-categories logit model instead of a cumulative logit model is not having to deal with order constraints on intercept parameters, which makes specification of prior distributions simpler.

We illustrate with the ulcer data of Table 11.4 that was analyzed in Section 11.3.4 using a cumulative logit model. For the analysis with a cumulative logit model, β was an assumed common cumulative log odds ratio. For the adjacent-categories logit model,

$$\log \frac{P(Y_i = j)}{P(Y = j + 1)} = \alpha_j - \beta x_i,$$

the corresponding loglinear model for the cell expected frequencies $\{\mu_{ij}\}$ is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta x_i v_j,$$

---

[1] For example, in version 9.2, SAS PROC GENMOD can handle Poisson or binomial responses but not multinomial.

TABLE 11.6. Posterior Estimates of Local Log Odds Ratio β for Adjacent-Categories Logit Analyses of Ulcer Data of Table 11.4

Prior Distribution	Mean	Std. Dev.	95% Credible Interval
Improper	0.878	0.335	(0.252, 1.566)
Normal ($\sigma = 1000$)	0.879	0.335	(0.251, 1.568)
Normal ($\sigma = 1.0$)	0.737	0.272	(0.218, 1.286)
Jeffreys	0.842	0.328	(0.229, 1.516)
ML[a]	0.842	0.328	(0.234, 1.534)

[a]Results in ML row are the ML estimate, standard error, and 95% profile likelihood confidence interval.

with unit-spaced scores such as $\{v_1 = -1, v_2 = 0, v_3 = 1\}$. This loglinear model is a special case of the linear-by-linear association model (6.2), and the parameter β is an assumed common local log odds ratio.

In Table 11.6 we summarize the ML frequentist analysis and Bayesian analyses[2] for various prior distributions for β and for the loglinear main effect parameters, with the ML results used for the initial values. The improper prior distribution is uniform over the entire parameter space. The normal prior distributions have means of 0. The standard deviation of 1000 corresponds to a highly diffuse prior similar to the improper uniform prior, whereas the standard deviation of 1.0 is relatively quite informative. The MCMC analysis was run sufficiently long (two million iterations) that the Monte Carlo standard errors for these results are less than 0.002 for each analysis. Results are substantively similar with each approach, with the informative prior resulting in some shrinkage of the estimated effect toward the prior mean.

11.3.6 Bayesian Fitting of Continuation-Ratio Logit Models

In Section 4.3 we observed that ML fitting of continuation-ratio logit models can utilize the connection between a multinomial likelihood and a product binomial likelihood. This results in two simplifications for Bayesian fitting of such models: First, Bayesian software that can handle binomial responses can be used for Bayesian fitting of such models. Second, the model does not have an ordering constraint on the cutpoint parameters.

We illustrate using continuation-ratio logit modeling of Table 4.3, analyzed by frequentist methods with such a model in Section 4.2.6. The table cross classifies a sample of children by their tonsil size, with categories (not enlarged, enlarged, greatly enlarged), and by whether they were carriers of a bacteria that is the cause of streptococcal infections.

Let x indicate whether a child is a carrier of the bacteria (yes = 0.50, no = -0.50). The sequential model of proportional odds form is

$$\log \frac{\pi_1(x)}{\pi_2(x) + \pi_3(x)} = \alpha_1 + \beta x, \qquad \log \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta x.$$

[2]Obtained using PROC GENMOD in SAS.

TABLE 11.7. Posterior Estimates of Effect β for Continuation-Ratio Logit Analyses of Enlarged Tonsils Data of Table 4.3

Prior Distribution	Mean	Std. Dev.	95% Credible Interval
Improper	−0.532	0.199	(−0.926, −0.146)
Normal ($\sigma = 1000$)	−0.533	0.199	(−0.926, −0.146)
Normal ($\sigma = 1.0$)	−0.518	0.194	(−0.902, −0.141)
Jeffreys	−0.528	0.198	(−0.919, −0.144)
ML[a]	−0.5285	0.198	(−0.922, −0.144)

[a]Results in ML row are the ML estimate, standard error, and 95% profile likelihood confidence interval.

Table 11.7 summarizes the ML frequentist analysis and Bayesian analyses[3] for various prior distributions with the ML results used for initial values. The normal prior distributions have means of 0. The MCMC analysis was run sufficiently long (two million iterations) that the Monte Carlo standard errors for these results are less than 0.002 for each analysis. Results are substantively similar with each approach, with the informative prior ($\sigma = 1.0$) resulting in some shrinkage of the estimated effect toward the prior mean.

11.4   BAYESIAN ORDINAL ASSOCIATION MODELING

Bayesian modeling of association between two ordinal variables provides an alternative to the frequentist association modeling of Chapter 6. With bivariate ordinal responses summarized by contingency tables having ordered rows and ordered columns, we analyze the degree of evidence supporting an association in a particular direction, positive or negative, and estimate its strength.

11.4.1   Evaluating Stochastic Ordering for Two Ordered Multinomials

When one variable (say, the row variable) has only two categories, it is often useful to compare the conditional distributions on the columns for the two rows. This is true even if the variables are both response variables rather than one being an explanatory variable. We then can apply the Bayesian modeling methods of Section 11.3, regarding the two rows as groups to be compared with a model such as the cumulative link model $h[P(Y_i \leq j)] = \alpha_j - \beta x_i$, applied with an indicator predictor for comparing two multinomial distributions.

An alternative approach is to summarize the degree of evidence supporting a stochastic ordering of the two rows with respect to their conditional distributions on the column variable. The cumulative link model implies that the two distributions are stochastically ordered, but this section shows an alternative analysis that does not assume a particular structural model or even assume a stochastic ordering.

[3]Obtained using PROC GENMOD in SAS, as shown in Table A.3 in the Appendix.

Let $\pi_{j|i}$ denote the probability of outcome j in group i, $i = 1, 2$. Recall that group 2 is stochastically larger than group 1 if

$$\sum_{j=1}^{k} \pi_{j|2} \le \sum_{j=1}^{k} \pi_{j|1}, \qquad k = 1, 2, \ldots, c - 1.$$

Equivalently, all $c - 1$ cumulative log odds ratios are nonnegative. For $c = 2$, this is merely the ordering of the success probabilities, $\pi_{1|2} \le \pi_{1|1}$. In that case, one group is necessarily stochastically larger than the other. When $c > 2$, the groups need not be stochastically ordered, because one set of cumulative probabilities need not be bounded above by the other set for all j.

Altham (1969) used a Bayesian analysis with two ordinal multinomial distributions that evaluates the extent of evidence about stochastic ordering. Assuming independent Dirichlet prior distributions, she obtained an expression for the posterior probability that one distribution is stochastically larger than the other, that is, that each cumulative probability for one distribution is no greater than the corresponding cumulative probability for the other distribution.

Suppose that the Dirichlet prior distributions for $\{\pi_{j|1}\}$ and $\{\pi_{j|2}\}$ have prior parameters $\{\alpha_j > 0\}$ for $\{\pi_{j|1}\}$ and $\{\beta_j > 0\}$ for $\{\pi_{j|2}\}$. Denote the cell counts in row i by $\{n_{i1}, \ldots, n_{ic}\}$. Assuming $\{n_{1j}\}$ and $\{n_{2j}\}$ are independent multinomial samples, the posterior distributions are independent Dirichlet with parameters $\{\mu_j = n_{1j} + \alpha_j\}$ and $\{\nu_j = n_{2j} + \beta_j\}$. Let $\mu = (\mu_1 + \cdots + \mu_c)$ and $\nu = (\nu_1 + \cdots + \nu_c)$ be the effective posterior sample sizes. Based on the equivalence of a Dirichlet random variable with differences between certain order statistics from a uniform distribution over $[0, 1]$, Altham (1969) found the posterior probability that group 2 is stochastically larger than group 1. It equals

$$\sum_{s_1} \cdots \sum_{s_c} \frac{\binom{\mu_1 + \nu_1 - 1}{s_1}\binom{\mu_2 + \nu_2}{s_2} \cdots \binom{\mu_{c-1} + \nu_{c-1}}{s_{c-1}}\binom{\mu_c + \nu_c - 1}{s_c}}{\binom{\mu + \nu - 2}{\nu - 1}},$$

where each s_j index varies between 0 and the upper limit in the corresponding binomial coefficient, but such that $(s_1 + \cdots + s_j) \le (\mu_1 + \cdots + \mu_j - 1)$ for $1 \le j \le c - 1$. For fixed (μ, ν) and fixed $(\mu_j + \nu_j)$ for each j, this posterior probability is monotone increasing as $(\mu_1 + \cdots + \mu_j)$ increases for any $j < c$, that is, as relatively more posterior probability falls at the low end of the scale for group 1. The nearer this posterior probability is to 1, the stronger the evidence that group 2 tends to make higher responses than the group 1.

When $c > 2$, the posterior probability that group 1 is stochastically larger than group 2 is not the complement of the posterior probability that group 2 is stochastically larger than group 1, because the groups need not be stochastically ordered. For example, in a 2×4 table with posterior Dirichlet parameters $(\mu_1, \mu_2, \mu_3, \mu_4) = (3, 4, 10, 11)$ in row 1 and $(\nu_1, \nu_2, \nu_3, \nu_4) = (2, 6, 9, 11)$ in row 2, Altham calculated that the probability is 0.17 that group 1 is stochastically larger than group 2 and 0.23 that group 2 is stochastically larger than group 1.

11.4.2 Example: Comparing Operations for Ulcer Revisited

Weisberg (1972) presented an algorithm for performing Altham's test about stochastic ordering, illustrating the method for the data in Table 11.4 comparing two operations for duodenal ulcer. Weisberg used uniform Dirichlet priors with $\alpha_1 = \alpha_2 = \alpha_3 = 1$ and $\beta_1 = \beta_2 = \beta_3 = 1$, corresponding to little prior information about the probabilities. The posterior probability is 0.975 that the response distribution is stochastically larger (i.e., better) for operation 2 than for operation 1.

Although the posterior probability calculated in this analysis takes into account the category ordering, the basic model itself does not do so. That is, the Dirichlet prior distribution is invariant to permutation of the categories, and no relation is specified between the probabilities in the two rows. This analysis is an alternative to formulating a standard ordinal model for the data, such as a cumulative logit model, and finding the posterior probability that the parameter comparing operation 2 to operation 1 is positive. Assuming such a model, this posterior probability is the complement of the posterior probability that the parameter is negative. In Section 11.3.4 we performed such an analysis and obtained similar substantive conclusions.

Yet another possible analysis for a table of this sort is Bayesian inference for a summary measure of association for $2 \times c$ tables such as the measures of stochastic superiority α and Δ. Given the cell counts $\mathbf{n}$, finding a posterior probability such as $P(\Delta > 0 \mid \mathbf{n})$ provides a Bayesian analog of the frequentist Wilcoxon test based on midrank scores. One way to do this generalizes the analysis Altham (1969) proposed for 2×2 tables. Independent Dirichlet prior distributions for the conditional probabilities in the two rows combined with independent multinomial likelihood functions yield independent posterior Dirichlet distributions for those conditional probabilities. Those posterior distributions induce a posterior distribution for Δ and α.

11.4.3 Modeling Ordinal Association in Contingency Tables

For $r \times c$ tables, simple but useful models for describing ordinal association include the uniform local odds ratio model introduced in Section 6.6 and the uniform global odds ratio model introduced in Section 6.5. The uniform local odds ratio model is a special case of the model (6.2) of linear-by-linear association. For fixed monotone row scores $\{u_i\}$ and monotone column scores $\{v_j\}$, that model for the expected frequencies $\{\mu_{ij}\}$ in a two-way contingency table is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j.$$

With unit-spaced scores, β is the uniform value for the local log odds ratios. The sign of β is equivalent to the sign of the correlation for the joint distribution. Given the cell counts $\mathbf{n}$, the posterior $P(\beta > 0 \mid \mathbf{n})$ or its complement summarizes the evidence about the direction of the association. A posterior credible interval for β describes the strength of association.

To illustrate, consider again the data in Table 11.4 from a study comparing two operations for duodenal ulcer. In Section 11.3.5 we conducted a Bayesian analysis for this model as a way of conducting Bayesian inference for an adjacent-categories logit model. The results shown in Table 11.6 for that logit model also apply to the uniform local log odds ratio for the linear-by-linear association model with unit-spaced scores.

Section 6.5.2 presented a generalized association model, the RC model (6.13), in which the scores are themselves unknown parameters. Evans (1993) provided a Bayesian analysis for this model. Based on independent normal priors with large variances for the loglinear parameters for the saturated model, they used a posterior distribution for loglinear parameters to induce a marginal posterior distribution on the RC submodel through Euclidean projection. The RC model treats the variables as nominal. To treat them, instead, as ordinal, Iliopoulos et al. (2007, 2009) imposed an ordering constraint on both sets of score parameters. Their Bayesian solutions are alternatives to the frequentist analysis of Section 6.5.6. The 2007 article used gamma priors for differences between adjacent scores that are constrained to be positive, whereas the 2009 article treated the scores as ordered uniform variables and focused on model comparison, using a MCMC algorithm that allowed moving between parameter spaces of different sizes. For the simpler row effects model (6.5), Tarantola et al. (2008) proposed a Bayesian approach for analyzing whether the rows can be grouped into clusters, with all rows in the same cluster having the same row effects.

The RC model extends to the more general RC(M^*) model introduced in Section 6.5.10. It has M^* sets of row and column parameters, with $M^* \leq \min(r - 1, c - 1)$. Kateri et al. (2005) provided a Bayesian analysis and addressed the issue of determining the order of association M^*. They checked the fit by evaluating the posterior distribution of the distance of the model from the full model.

11.4.4 Mental Health and SES Revisited

In Section 6.5.3 we analyzed a 6×4 contingency table (Table 6.6) relating $Y =$ child's mental health status (well, mild symptoms, moderate symptoms, impaired) with $X =$ parents' socioeconomic status (row 1 = highest to row 6 = lowest). The uniform local odds ratio version of the linear-by-linear association model fits very well (deviance of 9.9 with df $= 14$), with $\hat{\beta} = 0.091$ (SE $= 0.015$). There is very strong evidence that mental health tends to be better at higher levels of parents' SES. With independent normal priors with means of 0, a Bayesian analysis produces a 95% posterior credible interval for β of $(0.062, 0.120)$ when each $\sigma = 1000$ and $(0.060, 0.119)$ when each $\sigma = 1.0$. This is consistent with the frequentist analysis, not surprising as the sample size ($n = 1660$) was large. With such a large sample size, it takes a very sharp prior distribution to have much of a smoothing influence.

Albert (1997) showed how to compare the uniform association model to the more general loglinear model with association parameters $\{\lambda_{ij}^{XY}\}$. He used a vague uniform prior for β and independent normal priors with constant variance for the set of association parameters. He also considered an alternative model that satisfies

independence except for some outlier cells. His approach yields estimates of the association parameters that allow for uncertainty that the hypothesized model is correct. He concluded that these data were equally well represented by an outlier model and by the uniform association model. His outlier model allowing three outliers provided estimates of log local odds ratios that are quite close to 0 relative to their standard errors, the largest one being the local log odds ratio estimate of 0.28 (SE = 0.23) at the two lowest levels of SES and two highest levels of mental health status.

For their Bayesian approach with the RC model for these data, Evans et al. (1993) obtained similar results as Section 6.5.3 reported for the RC-model frequentist analysis. For example, they found strong evidence of a positive association but with the first two scores for SES having posterior means slightly out of order.

To estimate the cell probabilities themselves, we could use one of these model fits or we could use the empirical Bayesian method of Section 11.2.4 to smooth the sample proportion estimates $\{p_{ij}\}$ toward the ML fit of the model. Agresti and Chuang (1989) used a Dirichlet prior distribution for $\{\pi_{ij}\}$ with Dirichlet means $\{\gamma_{ij}\}$ that are themselves estimated by the ML fitted probabilities for the model of uniform local odds ratios. For these data, this provides Bayesian estimators of the form

$$0.28 p_{ij} + 0.72 \hat{\gamma}_{ij}.$$

Even though the sample size is large ($n = 1660$), these give primary weight to the ML model fit because the fit is so good.

11.5 BAYESIAN ORDINAL MULTIVARIATE REGRESSION MODELING

In modeling multivariate ordinal response variables, the focus can be on the association and interaction structure among the variables, extending methods of Section 11.4.3. More commonly, in practice in longitudinal studies and other types of studies having clustered observations, the focus is on the regression modeling of each component of the response vector in terms of explanatory variables, while accounting for the clustered nature of the ordinal responses. In Chapters 8 to 10 we presented corresponding frequentist methods.

11.5.1 Modeling Multivariate and Hierarchical Responses

In Chapter 10.1 we analyzed multivariate ordinal response data using random effects models. That approach is somewhat Bayesian in flavor, as it specifies probability distributions for the random effect terms in the model. With a fully Bayesian approach, no effects are fixed, in the sense that all effects are given prior distributions. Nonetheless, it is helpful in building a model to distinguish between terms for the clusters, such as a random intercept for subjects in a longitudinal study, and terms for the effects of explanatory variables.

Best et al. (1996) used Bayesian methods with a cumulative logit model of proportional odds form having a random intercept, as introduced in Section 10.1.1. They presented their modeling in the form of a graphical model that makes explicit the conditional independence structure.

The majority of work for multivariate ordinal responses, however, has focused on multivariate probit models. This is because such models naturally relate to latent variable models for multivariate normal responses. Chib and Greenberg (1998) showed this for the binary case in extending the Albert and Chib (1993) approach for a univariate response. A multivariate normal latent random vector with cut-points along the real line defines the categories of the observed discrete variables. The correlation among the categorical responses is induced through the covari-ance matrix for the underlying latent variables. The authors' approach includes a method for sampling the posterior distribution of the correlation matrix. Chen and Dey (2000) used a similar latent structure but focused on ordinal responses and (as in Albert and Chib 1993) permitted greater flexibility through using scale mix-tures of normals for the latent variables, thus encompassing other link functions including a t link that is a close approximation to the logit. They also suggested a residual analysis on the latent scale.

Chen and Dey illustrated their methodology with an item response example having a five-category ordinal response scale. The data resulted from a survey of secondary school teachers who were asked about the importance of various features of Master's degree programs for mathematics teachers. They noted that the exchangeability assumption implicit in ordinary frequentist random intercept models is often invalid. An advantage of their approach is that the correlation need not have an exchangeable structure within clusters of observations.

11.5.2   Other Research on Ordinal Multivariate Models

In another variation of Bayesian inference for an ordinal multivariate probit model with an underlying latent structure, Kottas et al. (2005) used a nonparametric Bayesian approach with a Dirichlet mixture of multivariate normals for the prior distribution. They showed that using a mixture of normals allows the dependence structure to vary across the contingency table cross-classifying the responses. They illustrated with an example on interrater agreement, suggesting that raters may tend to agree about extremely high or low scores but show more disagreement in rating average observations. They noted also that inference in some ordinal multivari-ate probit models is plagued by problems related to the choice and resampling of cutpoints defined for the latent variables, because the cutpoints tend to be highly correlated with the latent variables. They claimed that this problem does not occur with their approach, because without loss of generality their approach treats the cutpoints as fixed.

Webb and Forster (2008) parameterized the multivariate ordinal probit model in such a way that conditional posterior distributions are standard and easily simulated. They compared different conditional independence specifications by obtaining pos-terior distributions over classes of graphical models, comparing posterior marginal probabilities of the models given the data (integrating out the parameters).

For other Bayesian research that has used an underlying latent random vector to induce an ordinal multivariate cumulative link model, see Lunn et al. (2001) and Congdon (2005, Sec. 6). In Note 11.7 we summarize other Bayesian research on multivariate ordinal models.

11.6   BAYESIAN VERSUS FREQUENTIST APPROACHES TO ANALYZING ORDINAL DATA

We have not discussed philosophical issues underlying the choice between the frequentist and Bayesian frameworks for analyzing data, as that has received considerable attention in the Bayesian literature. Our comments here pertain solely to practical aspects of using the two approaches.

The frequentist approach to analyzing ordinal data has the advantage of relative simplicity. We need only to specify the model and the distribution of the data, which determine the likelihood function. The Bayesian approach has two further complications: First, it is necessary to specify a prior distribution. We've seen that this can be nontrivial even for a relatively simple model for a univariate ordinal response. Second, it can be rather complicated to compute adequately (to proper convergence) the posterior distribution and its summaries.

The choice of the prior distribution and appropriate specification of a model requires careful thought. For example, in the cumulative logit model example of Section 11.3.4, results depend strongly on the choice of prior distribution and on the way of defining an indicator variable for an explanatory variable. For multivariate models, it can be quite challenging to determine a sensible prior distribution. An independent observer may worry that results are not sufficiently objective, perhaps being too strongly influenced by the choice of prior distribution.

To attempt to reduce the subjective aspect of a Bayesian analyis, the data analyst may decide to use prior distributions that are flat relative to the likelihood function. But the implications of such choices can also be worrisome with complex models that have large numbers of parameters. Some highly influential statisticians are skeptical about Bayesian methods being used appropriately in such cases. For example, in his book *Principles of Statistical Inference* (Cambridge University Press, 2006), D. R. Cox states (p. 76): "Many empirical applications of Bayesian techniques use proper prior distributions that appear flat relative to the information supplied by the data. If the parameter space to which the prior is attached is of relatively high dimension the results must to some extent be suspect."

Both the choice of prior distribution and the proper computational implementation of the Bayesian approach require a fair amount of methodological and computational savvy. For all but the simplest models, the use of Bayesian methods should be considered rather daunting for the unsophisticated data analyst. Over time, as data analysts gain more experience with the Bayesian approach and as software becomes further developed and certain cases become considered as defaults when there is little prior information, these issues may be less problematic.

In the 1960s, many statisticians discovered the value of Bayesian analyses because of the advantages of employing shrinkage estimation. An example was

Charles Stein's classic result that the ordinary maximum likelihood estimator of a vector of means of normal distributions was dominated by a Bayes estimator. These days, it is possible to obtain the same advantages of shrinkage in a frequentist context, for example by using models with random effects. In this sense, the lines between Bayesian and frequentist analysis have blurred somewhat. Nonetheless, there are still some analysis aspects for which the Bayesian approach is a more natural one. One is providing a natural way to combine available prior information and new data. Another is the availability of model averaging to deal with the thorny issue of model uncertainty. Another is a common approach for conducting inference whether the sample size is large or small. In addition, to many methodologists it is more natural to make probability statements about unknown parameters than to make probability statements about hypothetical observations conditional on particular parameter values.

In recent years, relatively fewer statisticians take the dogmatic view of regarding only the frequentist approach or only the Bayesian approach as valid. In the future, it seems likely that most data analysts will feel comfortable using both paradigms, making their choice for a data analysis according to which approach seems more natural for the particular application.

CHAPTER NOTES

Section 11.2: Estimating Multinomial Parameters

11.1. For an ordinal response, Vijn (1983) expressed the Dirichlet density in terms of cumulative probabilities and in terms of cutpoints for an underlying latent variable. Sedransk (1985) and Gelfand (1992) estimated multinomial probabilities under the umbrella constraint $\pi_1 \leq ... \leq \pi_k \geq \pi_{k+1} \geq ... \geq \pi_c$, using a truncated Dirichlet prior and possibly a prior on k if it is unknown.

11.2. Frequentist methods also exist for smoothing contingency tables and shrinking sample data toward models. These include kernel smoothing, mazimizing penalized likelihoods, and generalized additive models and other semiparametric regression models containing smooth functions of explanatory variables as the predictors. Kernel smoothing estimates the mean at a particular point by a weighted averaging of data near that point, with less weight given to data further away. For ordinal data, see Brown and Rundell (1985) and Dong and Simonoff (1995). Penalized likelihood methods subtract a penalty term from the log-likelihood function, the penalty being smaller when the data are smoother. Simonoff (1987) argued that this method is superior to kernel methods. Penalized likelihood methods share some features with Bayesian fitting, as the penalty function results in the same type of smoothing as does imposition of a prior distribution, and the penalty function can be related to a prior distribution. For smoothing methods for ordinal data, see Simonoff (1996, Chap. 6; 1998) for surveys, Dong and Simonoff (1995), Hastie and Tibshirani (1987), Simonoff (1987), Titterington and Bowman (1985), Yee and Wild (1996), and Congdon (2005, Sec. 7.3). For the use of smooth predictor components in ordinal models using cumulative logits or continuation-ratio logits, see

Hastie and Tibshirani (1987), Yee and Wild (1996), Fahrmeir and Tutz (2001), Kauermann and Tutz (2003), and Tutz (2003). For discussion of ordinal regression pertaining to machine learning, see Herbrich et al. (1999), Frank and Hall (2001), Shashua and Levin (2003), Chu and Ghahramani (2005), Chu and Keerthi (2007), and Waegemana et al. (2008).

Section 11.3: Bayesian Ordinal Regression Modeling

11.3. Section 11.3 described regression models that assume a particular link function. For binary and ordinal regression, Lang (1999) used a parametric link function based on smooth mixtures of two extreme value distributions and a logistic distribution. His model used a flat, noninformative prior distribution for the regression parameters and was designed for applications in which some prior information exists about the appropriate link function. Ntzoufras (2009, pp. 235–236) summarized related research.

11.4. Section 5.4 presented a generalization of the cumulative link model that allows dispersion as well as location effects. Congdon (2005, Sec. 7.4) considered Bayesian inference for such a model. For other Bayesian ordinal regression analyses, see Albert and Chib (1993, 2001), Cowles et al. (1996), Bradlow and Zaslavsky (1999), Ishwaran and Gatsonis (2000), Ishwaran (2000), Xie et al. (2000), Congdon (2005, Chap. 7), and Gill and Casella (2009).

Section 11.4: Bayesian Ordinal Association Modeling

11.5. Bhattacharya and Nandram (1996) and Evans et al. (1997) estimated several multinomial distributions under a stochastic ordering when there is uncertainty about whether such a restriction is valid. Gelfand and Kuo (1991) modeled stochastically ordered outcomes for responses at different dosage levels in a bioassay.

11.6. Several articles have used Bayesian methods for modeling ordinal agreement. Strong association does not necessarily imply strong agreement, so association models alone are not usually adequate. Johnson (1996) proposed a Bayesian model for studies in which several judges provide ordinal ratings of items, a particular application being test grading. Johnson assumed that for a given item, a normal latent variable underlies the categorical rating. For a given judge, cutpoints define boundary points for the categories. He suggested uniform prior distributions over the real line for the cutpoints, truncated by their ordering constraints. The model is used to regress the latent variables for the items on covariates in order to compare the performance of raters. The model structure is hierarchical, with independent normal priors for the different judges but with the variances of those priors having an inverse gamma distribution. Kottas et al. (2005) also used a multivariate probit model but with a nonparametric prior structure. Mwalili et al. (2004) presented a model to correct for interobserver measurement error. Broemeling (2009) focused on Bayesian inference for agreement measures such as weighted kappa.

Section 11.5: Bayesian Ordinal Multivariate Regression Modeling

11.7. Lawrence et al. (2008) used MCMC methods with multivariate probit models by sampling correlation matrices using Gibbs sampling. Other examples of Bayesian approaches to ordinal modeling of multivariate or hierarchical data include Tan et al. (1999), Rossi et al. (2001), Biswas and Das (2002), Qiu et al. (2002), Kaciroti et al. (2006), and the references in Note 10.5. For modeling of case–control ordinal data, see Mukherjee et al. (2007).

EXERCISES

11.1. Outline how you would conduct a Bayesian analysis for an ordinal measure of association that is not necessarily connected with a model, such as (a) gamma and Kendall's tau-b, and (b) a global odds ratio.

11.2. For an ordinal $2 \times c$ contingency table, develop a Bayesian analog of the Wilcoxon test and Bayesian inference for the related measures of stochastic superiority, Δ and α. For Table 11.4, find a posterior tail probability using one of these measures that is analogous to the frequentist exact Wilcoxon one-sided P-value (which equals 0.0058) for the alternative that operation 2 is better than operation 1.

11.3. Refer to the example on government spending on the arts in Section 11.2.2. The corresponding counts for the sampled males in the 18-21 age range were (6, 4, 10, 5, 2). Using methods of this chapter, (a) report Bayesian estimates of the corresponding cell probabilities, (b) report Bayesian estimates for the table cross-classifying opinion by gender.

11.4. Try to replicate the results shown for modeling the ulcer data in Sections 11.3.4 and 11.3.5. Obtain results for the corresponding continuation-ratio logit model for these data.

11.5. Refer to the opinions about teen sex, premarital sex, and extramarital sex shown in Agresti and Lang (1993a) and Congdon (2005, p. 262). Analyze these data with (a) a marginal model, (b) a random effects model, and (c) a Bayesian model. Compare results and interpretations.

APPENDIX

Software for Analyzing Ordinal Categorical Data

All major statistical software has procedures for categorical data analyses. In this appendix we discuss the use of SAS, R, Stata, and SPSS, with brief summaries of other software. We do not attempt to provide detailed instructions, as information on that is available in specialized publications and at Internet sites that we list. Our goal is merely to provide information about procedures that are available for the ordinal analyses presented in this book. Many of the Internet addresses listed below will be out of date at some stage, but the reader should be able to find similar information with a search on the Internet using relevant keywords. For more information about software packages for categorical data analyses, see the web site www.stat.ufl.edu/~aa/cda/software.html.

SAS

In SAS, the main procedures (PROCs) for categorical data analyses are FREQ, GENMOD, LOGISTIC, and NLMIXED. PROC FREQ provides large- and small-sample tests of independence in two-way tables, measures of association and their estimated standard errors, and generalized CMH tests of conditional independence. PROC GENMOD fits cumulative link models and loglinear models for ordinal responses, and it can perform GEE analyses for marginal models as well as Bayesian model fitting for some cases. PROC LOGISTIC also fits cumulative link models. PROC NLMIXED fits models with random effects and generalized nonlinear models. PROC CATMOD can fit a wide variety of models, mainly using WLS but with ML for models that can be expressed using baseline-category logits, such as adjacent-categories logit models. The examples below show SAS code (version 9.2) for many ordinal analyses. For convenience, examples enter data in the form of the contingency table displayed in the text. In practice, data would usually be listed

at the subject level. For further information about using SAS for ordinal categorical data analyses, see Stokes et al. (2000), Bender and Benner (2000), the appendix of O'Connell (2006), and the web site www.ats.ucla.edu/stat/sas/examples/icda at UCLA.

PROC GENMOD can fit cumulative link models by specifying DIST = MULTINOMIAL and LINK = CLOGIT (cumulative logit) or LINK = CPROBIT (cumulative probit) or LINK = CCLL (cumulative complementary log-log).

PROC LOGISTIC can also fit cumulative link models and conduct the score test of the proportional odds assumption of identical effect parameters for each cutpoint. The first three PROC statements in Table A.1 fit cumulative logit models to Table 3.1. Stokes et al. (2000) showed how to use PROC GENMOD with the GEE methodology to obtain estimates for a partial proportional odds model.

Adjacent-categories logit models can be fitted in SAS with PROC CATMOD by fitting equivalent baseline-category logit models. Table A.2 fits the adjacent-categories-logit model to Table 4.1. PROC CATMOD has options (CLOGITS and ALOGITS) for fitting cumulative logit and adjacent-categories logit models to ordinal responses; however, those options provide weighted least squares (WLS) rather than ML fits and should be used only with nonsparse contingency tables. CATMOD treats zero counts as structural zeros, so they must be replaced by small constants (such as 10^{-8}) when they are actually sampling zeros. PROC CATMOD can also fit mean response models (Section 5.6) using WLS. Continuation-ratio

TABLE A.1. SAS Code for Ordinal Modeling of Table 3.1

```
data astrosci;
input degree astro count v2; uv=astro*degree; uv2=v2*degree;
     datalines;
         0 1 23 1
         0 2 84 3
         0 3 98 4
         1 1 50 1
...
         4 2 23 3
         4 3 148 4
         ;
proc genmod; weight count; * cumulative logit;
 model astro = degree / dist=multinomial link=clogit lrci type3;
proc logistic; weight count; * cumulative logit;
 model astro = degree / aggregate scale=none;
proc genmod; weight count; class degree; * treat degree as nominal;
 model astro = degree / dist=multinomial link=clogit lrci type3;
proc genmod; class astro degree; * linear-by-linear assoc;
 model count = astro degree uv / dist=poi link=log type3 obstats;
proc genmod; class astro degree; * linear-by-linear assoc, other
 scores;
 model count = astro degree uv2 / dist=poi link=log type3 obstats;
proc freq; weight count; * generalized CMH tests;
 tables astro*degree / cmh;
```

TABLE A.2. SAS Code Showing Use of PROC CATMOD to Fit Adjacent-Categories Logit Model to Table 4.1

```
data stemcell;
input fund scresrch gender count;
     datalines;
     1 1 0 21
     1 1 1 34
     1 2 0 52
     1 2 1 67
     1 3 0 24
     1 3 1 30
     1 4 0 15
     1 4 1 25
     2 1 0 30
...
     3 4 1 12
        ;
proc catmod order=data; weight count;
population fund gender;
model scresrch = (1 0 0 3 0, 0 1 0 2 0, 0 0 1 1 0,
                  1 0 0 3 3, 0 1 0 2 2, 0 0 1 1 1,
                  1 0 0 6 0, 0 1 0 4 0, 0 0 1 2 0,
                  1 0 0 6 3, 0 1 0 4 2, 0 0 1 2 1,
                  1 0 0 9 0, 0 1 0 6 0, 0 0 1 3 0,
                  1 0 0 9 3, 0 1 0 6 2, 0 0 1 3 1) / ML NOGLS;
run;
```

logit models can be fitted using PROC GENMOD or PROC LOGISTIC by applying ordinary binary logistic regression models to the independent binomials to which the models apply, as explained in Section 4.2.6. Table A.3 illustrates for Table 4.3. Kuss (2006) showed how to use PROC NLMIXED to fit the stereotype model.

To fit the nonlinear location-scale model (5.4), Cox (1995) used an iteratively re-weighted Gauss–Newton algorithm, implemented with PROC NLIN in SAS. PROC NLMIXED could also be employed by providing in the code the likelihood function to be maximized.

The association models of Chapter 6 that are special cases of loglinear models, such as the linear-by-linear association model and the row effects model, can be fitted as special cases of generalized linear models using PROC GENMOD. As in ordinary loglinear modeling, one assumes a Poisson distribution for the cell counts and uses the log link function. The fourth and fifth PROC statements in Table A.1 fit the linear-by-linear association model (6.2) (with equally spaced scores and with column scores 1, 3, 4) to Table 6.1. The defined variable uv represents the cross-product of row and column scores, which has β parameter as coefficient in model (6.2). The RC model (6.13) can be fitted using PROC NLMIXED in SAS by specifying the likelihood function to be maximized, such as in Kuss (2006) for the related stereotype model. Davis (1988) showed how to fit the RC model using PROC MATRIX.

TABLE A.3. SAS Code for Frequentist and Bayesian Continuation-Ratio Logit Modeling of Table 4.3

```
data tonsils; * look at data as indep. binomials;
    input stratum carrier success failure;
    n = success + failure; carrier2 = carrier - 0.5;
datalines;
    1 1 19 53
    1 0 497 829
    2 1 29 24
    2 0 560 269
;
proc genmod data=tonsils; class stratum; * frequentist analysis;
 model success/n = stratum carrier / dist=binomial link=logit lrci
 type3;
run;
proc genmod data=tonsils; class stratum; * Bayesian analysis;
 model success/n = stratum carrier2 / dist=binomial link=logit;
 bayes coeffprior=normal (var=1.0) initialmle diagnostics=mcerror
 nmc=2000000;
run;
```

PROC FREQ estimates several measures of association and their standard errors (MEASURES option) and provides ordinal statistic (7.5) with a 'nonzero correlation' test (CMH1). The polychoric correlation is available with the PLCORR option in PROC FREQ. For tables having small cell counts, the EXACT statement can provide various exact analyses. These include exact trend tests for $r \times 2$ tables (TREND), and exact correlation tests for $r \times c$ tables (MHCHI). With the CMH option, PROC FREQ provides the generalized CMH tests of conditional independence presented in Section 6.4.5. The statistic for the "row mean scores differ" alternative treats X as nominal and Y as ordinal, and the statistic for the "nonzero correlation" alternative treats X and Y as ordinal. Table A.4 shows analyses of Table 2.3.

TABLE A.4. SAS Code for Measures of Association and Analyses of Data in Table 2.3

```
data gss;
input income happy count;
datalines;
  1 1 272
  1 2 294
  1 3 49
. . .
  3 3 208
      ;
proc freq data=gss; weight count;
 tables income*happy / chisq cmh cmh2 measures plcorr;
proc freq data=gss; weight count;
 tables income*happy / cmh cmh2 scores=rank measures;
```

The matched-pairs models of Chapter 8 that are special cases of loglinear or logistic models, such as the ordinal quasi-symmetry model (8.6) and ordinal agreement model (8.20), can be fitted as generalized linear models using PROC GENMOD. Table A.5 shows analyses of Table 8.2. The qi factor invokes the δ_i main-diagonal parameters in equation (8.10). It takes a separate level for each cell on the main diagonal, and a common value for all other cells. The *main-diag* term invokes the δ parameter in the ordinal agreement model. The bottom of Table A.5 fits logit models for the data entered in the form of pairs of cell counts (n_{ij}, n_{ji}). These three sets of binomial counts are labeled as *above* and *below* with reference to the main diagonal. The variable defined as *score* is the distance $(u_j - u_i) = j - i$. The model fitted without an intercept (NOINT option) is ordinal quasi-symmetry. For rater agreement (Section 8.5), the AGREE option in PROC FREQ in SAS provides Cohen's weighted kappa with SE values. It uses the weights $\{w_{ij} = 1 - |i - j|/(c - 1)\}$ by default and the weights $\{w_{ij} = 1 - (i - j)^2/(c - 1)^2\}$ when you specify (WT = FC) with the AGREE option.

Table A.6 uses PROC CATMOD to fit the adjacent-categories logit paired preference model (8.22) to the soft-drink tasting data of Table 8.9. Here the equivalent model is fitted using baseline-category logits, for which the first threshold parameter is 0 and there is a common threshold parameter for categories 2 and 4 and a separate one for category 3, as explained in Section 8.6.3.

PROC GENMOD in SAS can implement the GEE method presented in Chapter 9, using the REPEATED statement to specify the variable name that identifies the subjects for each cluster. For multinomial responses, independence is currently the only working correlation structure for GEE. For a SAS macro with other working correlation structures, see Williamson et al. (1999). The TYPE3 option with the GEE approach provides score-type tests about effects. See Stokes et al. (2000, Sec. 15.11) for the use of GEE with missing data. PROC NLMIXED extends GLMs to GLMMs by including random effects. Although the multinomial distribution is not supported directly by NLMIXED, one can define the general likelihood function needed to fit the models through the use of SAS programming statements. Table A.7 shows how to analyze Table 8.5. The text web site www.stat.ufl.edu/~aa/ordinal/data.html shows NLMIXED code for the cumulative logit random effects analysis of the multicenter clinical trial data of Section 10.3.4 and for an adjacent-categories logit random effects analysis of a 3^4 movie reviewer data set in Hartzel et al. (2001b). For other discussion of using SAS with clustered data, see Molenberghs and Verbeke (2005).

Bayesian analyses for generalized linear models are available with the BAYES statement in PROC GENMOD. As of version 9.2, this is not available for multinomial distributions, but it can be employed for binomial and Poisson models, so this procedure is useful for Bayesian implementation of continuation-ratio logit models through the binomial factorization described in Section 4.2 and for Bayesian implementation of adjacent-categories logit models through the Poisson loglinear connection. The default prior distribution is improper uniform, but normal and Jeffreys priors are also easily invoked. Table A.3 illustrates with standard normal prior distributions for continuation-ratio logit modeling of Table 4.3, using the ML

TABLE A.5. SAS Code Showing Square-Table Analyses of Table 8.2

```
data gss_2006;
input helphlth helpenv count symm qi above maindiag diag1 diag2;
assoc = helphlth*helpenv;
datalines;
    1 1 199 1 1 0 1 0 0
    1 2  81 2 4 1 0 1 0
    1 3  83 3 4 1 0 0 1
    2 1 129 2 4 0 0 0 0
    2 2 167 4 2 0 1 0 0
    2 3 112 5 4 1 0 1 0
    3 1 164 3 4 0 0 0 0
    3 2 169 5 4 0 0 0 0
    3 3 363 6 3 0 1 0 0
        ;
proc genmod; class symm;
 model count = symm helpenv helphlth / dist=poi link=log; * ord quasi
 symm;
proc genmod; class helphlth helpenv qi;
 model count = helpenv helphlth qi assoc / dist=poi link=log; * quasi
 unif assoc;
proc genmod; class helphlth helpenv;
 model count = helpenv helphlth assoc maindiag / dist=poi link=log; *
 ord agreement;
proc genmod; class symm;
 model count = symm above / dist=poi link=log; * cond symmetry;
proc genmod; class symm;
 model count = diag1 diag2 symm / dist=poi link=log; * diag para
 symmetry;

data square;
input score below above @@; trials = below + above;
datalines;
1 4 43  1 8 99  1 18 230  2 4 163  2 1 185  3 0 233
;
proc genmod data=square;
 model below/trials = score / dist=bin link=logit noint;
```

estimates for initial values and reporting the Monte Carlo error as a diagnostic. The default number of iterations after the burn-in is 10,000 but can be changed by setting NMC. In version 9.2, multinomial models can be fitted using PROC MCMC. Table A.8 shows this for the Bayesian cumulative logit analysis of the ulcer operation data presented in Section 11.3.4. For details about using PROC MCMC for such a model, see the "Bayesian multinomial model for ordinal data" example at http://support.sas.com/rnd/app/examples/.

R

A good, detailed discussion of the use of R (and Splus) for models for categorical data is available online in the free manual prepared by Laura Thompson

TABLE A.6. SAS Code for Paired Preference Analyses of Table 8.8 with Model (8.22) for Adjacent-Categories Logits

```
data coke;
input row coke classic outcome count;
datalines;
1 1 -1  1 12
1 1 -1  2 19
1 1 -1  3 11
1 1 -1  4 14
1 1 -1  5 5
2 1  0  1 11
2 1  0  2 18
2 1  0  3 12
2 1  0  4 13
2 1  0  5 7
3 0  1  1 7
3 0  1  2 12
3 0  1  3 14
3 0  1  4 16
3 0  1  5 12
;
proc catmod data=coke order=data; weight count;
    population row;
model outcome =
  (0 0 -4  4,  1 0 -3  3,  0 1 -2  2,  1 0 -1  1,
   0 0 -4  0,  1 0 -3  0,  0 1 -2  0,  1 0 -1  0,
   0 0  0 -4,  1 0  0 -3,  0 1  0 -2,  1 0  0 -1)
  (1 2 = 'threshold', 3 4 = 'treatments')/ covb ML NOGLS
   pred=freq;
```

to accompany Agresti (2002) chapter by chapter. A link to this manual is at www.stat.ufl.edu/~aa/cda/software.html. Most of the ordinal methods considered in this book are also discussed in that manual; you can find them by searching the manual with relevant keywords [see also Bender and Benner (2000)].

As well as ordinary functions readily available with R, Laura Thompson's manual and this appendix mention useful specialized R functions from various R libraries. One of the most useful and powerful functions is *mph.fit*, written by Joseph Lang at the University of Iowa (jblang@stat.uiowa.edu). It can fit *multinomial–Poisson homogeneous models* that have the very general form

$$L(\mu) = X\beta$$

for probabilities or expected frequencies μ, where L is a general link function. Lang (2005) introduced this general class of models and their ML model fitting. One special case is the generalized loglinear form (6.18), namely $C \log A\mu = X\beta$. This model form includes the ordinal logit models of Chapters 3 and 4 and the association models of Chapter 6 (e.g., with global odds ratios or local odds ratios) that have linear predictors, the models for matched pairs in Chapter 8, and the marginal ordinal logit models of Chapter 9. Another special case is the model of

TABLE A.7. SAS Code for GEE and Random Intercept Cumulative Logit Analyses of Crossover Data in Table 8.5

```
data crossover;
    input a b c count symm;
    datalines;
        1 1 1 6    1
        1 1 2 4    2
        1 1 3 5    3
....
        3 3 3 0    10
;
data crossover;
set crossover;
    case = _n_;
    q1=1; q2=0; q3=0; resp=a; treat = 1; output;
    q1=0; q2=1; q3=0; resp=b; treat = 2; output;
    q1=0; q2=0; q3=1; resp=c; treat = 3; output;
data crossover;
set crossover;
y1=0; y2=0; y3=0;
 if q1=1 and a=1 then y1=1; if q1=1 and a=2 then y2=1; if q1=1
    and a=3 then y3=1;
 if q2=1 and b=1 then y1=1; if q2=1 and b=2 then y2=1; if q2=1 and
    b=3 then y3=1;
 if q1=0 and q2=0 and c=1 then y1=1; if q1=0 and q2=0 and c=2 then
    y2=1;
 if q1=0 and q2=0 and c=3 then y3=1;
proc genmod data=crossover; class case; freq count; * GEE analysis;
    model resp = q1 q2 / dist=multinomial link=clogit;
    repeated subject=case / type=indep;
proc nlmixed qpoints=200 data=crossover; * cumul logit random
    effects;
    bounds i2 > 0;
    eta1 = i1 + q1*beta1 + q2*beta2 + u; eta2 = i1 + i2 + q1*beta1
    + q2*beta2 + u;
    p1 = exp(eta1)/(1 + exp(eta1));
    p2 = exp(eta2)/(1 + exp(eta2)) - exp(eta1)/(1 + exp(eta1));
    p3 = 1 - exp(eta2)/(1 + exp(eta2));
    ll = y1*log(p1) + y2*log(p2) + y3*log(p3); model y1 ~ gen-
eral(ll);
    random u ~ normal(0, sigma*sigma) subject=case;
    replicate count; predict p1 out=new; proc print data = new;
proc nlmixed qpoints=200 data=crossover; * ACL random effects;
    eta1 = i1 + i2 + 2*q1*beta1 + 2*q2*beta2 + 2*u;
    eta2 = i2 + q1*beta1 + q2*beta2 + u;
    p1 = exp(eta1)/(1 + exp(eta1) + exp(eta2));
    p2 = exp(eta2)/(1 + exp(eta1) + exp(eta2));
    p3 = 1/(1 + exp(eta1) + exp(eta2));
    ll = y1*log(p1) + y2*log(p2) + y3*log(p3); model y1 ~ gen-
eral(ll);
    random u ~ normal(0, sigma*sigma) subject=case;
    replicate count; predict u out=new; proc print data = new;
```

TABLE A.8. SAS Code Using PROC MCMC for Bayesian Cumulative Logit Modeling of Table 11.4

```
data ulcer;
      input y1 y2 y3 treat;
      datalines;
7  17  76  -0.5
1  10  89   0.5
  ;
ods graphics on;
  proc mcmc data=ulcer nbi=10000 nmc=1000000 thin=2 seed=1181
    propcov=quanew monitor=(beta or);
    array alpha[2]; array gamma[2];
    parms alpha1 alpha2 beta;
    prior beta ~ normal(0,var=1000**2);
    prior alpha1 ~ normal(-1.0,var=1000**2);
    prior alpha2 ~ normal(1.0,var=1000**2,lower=alpha1);
    mu = beta*treat;
    do j = 1 to 2;
        gamma[j] = logistic(alpha[j] - mu);
    end;
    eta1 = gamma1; eta2 = gamma2 - gamma1; eta3 = 1 - gamma2;
    llike = logmpdfmultinom(of y1-y3, of eta1-eta3);
    model dgeneral(llike);
    beginprior;
        or = exp(beta);
    endprior;
run;
ods graphics off;
```

generalized linear form $A\mu = X\beta$, which includes the mean response model of Section 5.6. See www.stat.ufl.edu/~aa/ordinal/ord.html for examples of mph.fit for the analysis of standardized residuals in Section 3.5.8, the partial proportional odds analysis in Section 3.6.5, the mean response model in Section 5.6.2, the global odds ratio model in Section 6.6.2, the marginal cumulative logit model in Section 8.4.3, and the paired preference modeling of Section 8.6.4.

Another powerful library, developed by Thomas Yee at the University of Auckland, is VGAM for vector generalized linear and additive models (www.stat.auckland.ac.nz/~yee/VGAM). For details about models that can be fitted,[1] see Yee and Wild (1996), Yee and Hastie (2003), Yee (2010), and the Thompson manual. The *vglm* function fits by Fisher scoring a wide variety of models. The format resembles that for the ordinary R function *glm* with the addition of a family function. Possible models include the cumulative logit model (family function *cumulative*) with proportional odds or partial proportional odds or nonproportional odds, cumulative link models (family function *cumulative*) with or without common effects for each cutpoint, adjacent-categories logit models

---

[1] See also cran.us.r-project.org/doc/Rnews/Rnews 2008-2.pdf.

(family function *acat*), and continuation-ratio logit models (family functions *cratio* and *sratio*). The *grc* function fits the RC model.

Many other R functions can also fit cumulative logit and other cumulative link models. Thompson's manual (p. 121) describes the *polr* function from the MASS library, for which the cumulative logit is the default. The syntax is simple, such as with response variable y and explanatory variable x,

```
library(MASS)
fit.cum <- polr(y ~ x, data=tab3.1, method=''probit'')
summary(fit.cum)
```

Thompson also described the *lrm* function in the design library, the *lcr* function in the ordinal library, and the *nordr* function in the gnlm library. Continuation-ratio logit models can be fitted by the *lrm* and *nordr* functions, as Thompson discusses, or by applying ordinary binary logistic regression models (e.g., with the *glm* function) to the independent binomials to which the models apply, as explained in Section 4.2.6.

The *glm* function in R performs generalized linear modeling. The association models of Chapter 6 that are special cases of loglinear models, such as the linear-by-linear association model, can be fitted with the *glm* function, assuming a Poisson distribution for the cell counts and using the log link function. The syntax has a simple form, such as with cell counts and with factors x and y,

```
fit.LL <- glm(count ~ x + y + u:v, data=tab6.1, family=poisson)

summary(fit.LL)
```

See Thompson's manual for detailed examples.

The stereotype model and other nonlinear models for categorical data such as the RC model of Section 6.5 can be fitted with the *gnm* function developed[2] by David Firth and Heather Turner. The Goodman RC model can also be fitted by the *grc* function in the VGAM library mentioned above.

Chapters 2 and 7 introduced ordinal measures of association. A program written by Euijung Ryu that constructs confidence intervals for α (and hence Δ), including using Lang's *mph.fit* function to find the score interval, is available at www.stat.ufl.edu/~aa/cda/software.html. An R function *polychor* in the polycor package written by John Fox computes the polychoric correlation and its standard error.[3] The function *rcorr.cens* in the Hmisc library can compute gamma and Somers' d and their standard errors.[4] An R function *ci.table* prepared by Joseph Lang (jblang@stat.uiowa.edu) can obtain confidence intervals for measures of association by inverting score and likelihood-ratio tests.

---

[2]See www2.warwick.ac.uk/fac/sci/statistics/staff/research/turner/gnm
[3]see rss.acs.unt.edu/Rdoc/library/polycor/html/polychor.html
[4]See http://rweb.stat.umn.edu/R/library/Hmisc/html/rcorr.cens.html

The matched-pairs models of Chapter 8 that are special cases of loglinear or logistic models, such as the ordinal quasi-symmetry model, can be fitted with the *glm* function. See Thompson's manual for examples.

Lang's *mph.fit* function can use ML to fit many marginal models by expressing them as special cases of the generalized loglinear model (6.18). The package geepack contains a function *ordgee* for ordinal GEE analyses. Parsons et al. (2009) described a function *repolr* for GEE analysis with the cumulative logit model of proportional odds form. These permit working correlations other than independence.

The package glmmAK contains a function *cumlogitRE* for using MCMC to fit cumulative logit models with random effects.[5] The web site for the text *Bayesian Computation with R* by Jim Albert shows examples of some Bayesian categorical data analyses.[6]

STATA

In Stata, the *ologit* program (www.stata.com/help.cgi?ologit) fits cumulative logit models. After fitting the model, running *brant* provides the Brant (1990) Wald tests of the proportional odds assumption. The *predict* command produces estimated category probabilities, and the *prvalue* command produces them at particular values of explanatory variables. The *oprobit* program (www.stata.com/help.cgi?oprobit) fits cumulative probit models. A module *omodel* (written by R. Wolfe and W. Gould) available from Stata fits cumulative logit and probit models and provides an approximate likelihood-ratio test of the proportional odds assumption. Other ways to fit cumulative link models are with the Stata module OGLM. Adjacent-categories logit models can be fitted by fitting the corresponding loglinear models (Chapter 6) as generalized linear models with the *glm* program, assuming a Poisson distribution for the cell counts and using the log link function. See examples of both of these at the UCLA web site listed below. Continuation-ratio logit models can be fitted with the *ocratio* module (www.stata.com/search.cgi?query=ocratio) written by R. Wolfe and with the *seqlogit* module written by M. Buis. The *oglm* module written by R. Williams can fit cumulative link models having dispersion as well as location effects. His *gologit2* module can fit the partial proportional odds cumulative logit model. See Williams (2009). The stereotype model can be fitted with the *slogit* program (www.stata.com/help.cgi?slogit).

The association models of Chapter 6 that are special cases of loglinear models, such as the linear-by-linear association model, can be fitted with the *glm* program (www.stata.com/help.cgi?glm), assuming a Poisson distribution for the cell counts and using the log link function. The *tabulate* program (www.stata.com/help.cgi?tabulate_twoway) can generate many of the ordinal measures of association of Chapter 7, such as gamma and Kendall's tau-b, and their standard errors.

---

[5]See bm2.genes.nig.ac.jp/RGM2/pkg.php?p=glmmAK
[6]See bayes.bgsu.edu/bcwr. Another survey of Bayesian inference using R is at cran.r-project.org/web/views/Bayesian.html.

The GLLAMM module for Stata (see www.gllamm.org) can fit a very wide variety of models, including cumulative logit models with random effects. For details, see www.stata.com/search.cgi?query=gllamm and Chapter 7 of Rabe-Hesketh and Skrondal (2008).

For further information about Stata, see *Handbook of Statistical Analyses Using Stata*, 4th ed., by S. Rabe-Hesketh and B. Everitt (CRC Press, 2006). For methods for ordinal data, see Chapter 5 of *Regression Models for Categorical Dependent Variables Using Stata*, 2nd ed. by J. S. Long and J. Freese (Stata Press, 2006). For other examples of categorical data analyses, see also the useful site www.ats.ucla.edu/stat/stata/examples/icda at UCLA.

SPSS

SPSS can fit some ordinal multinomial models. On the ANALYZE menu, choose the REGRESSION option and then the ORDINAL suboption to get the ORDINAL REGRESSION menu for fitting a cumulative link model. Clicking on *Options*, you can request the link function, including the logit for cumulative logit models or the probit or complementary log-log. Clicking on *Output*, you can request a test of parallelism (i.e., proportional odds for the logit link). The model (5.4) with dispersion effects is also available as a regular option on the ORDINAL regression menu, by adding a *scale* component. The GENLOG function in SPSS can fit adjacent-categories logit models.

The association models of Chapter 8.6 that are special cases of loglinear models, such as the linear-by-linear association model and the row effects model, can be fitted as generalized linear models, assuming a Poisson distribution for the cell counts and using the log link function. On the ANALYZE menu, select the GENERAL-IZED LINEAR MODELS option and the GENERALIZED LINEAR MODELS suboption. Select the cell count as the *Dependent Variable* and then the Poisson for the *Distribution* and the log for the *Link Function*. Click on the *Predictors* tab at the top of the dialog box and then enter quantitative variables as *Covariates* and categorical variables as *Factors*. Click on the *Model* tab at the top of the dialog box and enter these variables as main effects, and construct any interactions that you want in the model. Click on *OK* to run the model.

The DESCRIPTIVE STATISTICS option on the ANALYZE menu has a suboption called CROSSTABS, which provides several methods of Chapter 7 for contingency tables. After identifying the row and column variables in CROSSTABS, clicking on *Statistics* provides a wide variety of options, including measures of association such as gamma, Kendall's tau-b, and Somers' *d* and their standard errors. It also provides a test statistic for testing that the true measure equals zero, which is the ratio of the estimate to its null standard error (which only applies under independence). SPSS also has an advanced module for small-sample inference (called *SPSS Exact Tests*) that provides exact P-values for various tests in CROSSTABS and NPAR TESTS procedures, such as exact tests of independence for contingency tables with ordinal classifications.

The matched-pairs models of Chapter 8 that are special cases of loglinear or logistic models, such as the ordinal quasi-symmetry model, can be fitted as

generalized linear models. For loglinear models, select the cell count as the *Dependent Variable*, the Poisson for the *Distribution*, and the log for the *Link Function*. For logistic models, the *Dependent Variable* is the binary outcome of whether in the cell in row i and column j or in the cell in row j and column i, the binomial is the *Distribution* of the outcome, and the logit is the *Link Function*.

For GEE methods, on the ANALYZE menu, select the GENERALIZED LINEAR MODELS option and the GENERALIZED ESTIMATING EQUATIONS (GEE) suboption. On the GEE window, click on *Repeated* and select the form for the working correlation model, and click on *Type of Model* to specify that you want a model for an ordinal logistic response. You can use a logit or probit link for the model.

For some examples of SPSS for ordinal modeling, see the appendix of O'Connell (2006).

OTHER PROGRAMS

See www.stat.ufl.edu/~aa/cda/software.html for links to information about the programs listed below.

StatXact and LogXact

For certain analyses, specialized software has greater capability than the major packages. A good example is StatXact (Cytel Software, Cambridge, Massachusetts). It provides small-sample categorical data analyses that do not rely on large-sample theory, such as the ordinal tests presented in Section 7.6. It can perform tests for two-way tables using criteria such as the correlation or rank correlation. It can perform analogs for ordered categorical responses of basic nonparametric methods such as the Wilcoxon test for $2 \times c$ tables or the Kruskal–Wallis statistic or Jonkheere–Terpstra statistic for $r \times c$ tables. It can conduct exact tests of conditional independence in three-way tables that are small-sample analogs of generalized CMH tests.

Most small-sample methods rely on the approach of eliminating unknown nuisance parameters from small-sample distributions by conditioning on their sufficient statistics. Computations for such methods require special algorithms to generate all the possible tables having the sufficient statistics that are fixed by the method. For those cases in which computations are too time consuming, StatXact uses simulation of the conditional distribution to estimate precisely the exact P-value.

LogXact, a companion program available from Cytel Software, conducts small-sample analyses for logistic regression model parameters. For an ordinal response, such inference is available for adjacent-categories logit models and continuation-ratio logit models. The conditional method of eliminating nuisance parameters does not work for cumulative logit models, because the cumulative logit is not the "canonical link" for a multinomial distribution, and non-canonical links do not have reduced sufficient statistics.

BUGS

BUGS (Bayesian Inference Using Gibbs Sampling) is statistical software for Bayesian modeling implemented using Markov chain Monte Carlo methods. The BUGS project was developed by the MRC Biostatistics Unit at the University of Cambridge, UK (www.mrc-bsu.cam.ac.uk/bugs). WinBUGS runs under Microsoft Windows. A version that emulates this is available for Macs, and an open-source version runs with Linux.

SuperMix

The SuperMix program, distributed by Scientific Software International, is designed for ML fitting of generalized linear mixed models. It can fit cumulative link models with random effects for logit, probit, and complementary log-log link functions, using Gauss–Hermite quadrature. The capability includes multi-level models that are difficult or very slow to fit by ML with most other software.

Latent Gold

Latent Gold is a program developed by Statistical Innovations (Belmont, Massachusetts) for fitting finite mixture models such as latent class models (i.e., the latent variable is categorical) in a generalized linear modeling framework. It can handle ordinal response variables and can include random effects that are treated in a nonparametric method rather than assumed to have a normal distribution.

GoldMineR

The GoldMineR (Graphical Ordinal Logit Displays Based on Monotonic Regression) package developed by Statistical Innovations (Belmont, Massachusetts) can fit a variety of models for ordered categorical response variables, including models that treat monotone scores for the categories as fixed or as parameters. The models include adjacent-categories logit models and some multiplicative models, such as the RC model and stereotype model.

SUDAAN

SUDAAN provides analyses for categorical and continuous data from stratified multistage cluster designs. It has a facility (MULTILOG procedure) for GEE analyses of marginal models for nominal and ordinal responses.

Others

Other programs that can be useful for various ordinal modeling include the econometric software LIMDEP (www.limdep.com) and the LEM program by J. K. Vermunt for categorical data modeling (spitswww.uvt.nl/~vermunt).

Bibliography

Agresti, A. 1976. The effect of category choice on some ordinal measures of association. *J. Amer. Statist. Assoc.* **71**: 49–55.

Agresti, A. 1977. Considerations in measuring partial association for ordinal categorical data. *J. Amer. Statist. Assoc.* **72**: 37–45.

Agresti, A. 1980. Generalized odds ratios for ordinal data. *Biometrics* **36**: 59–67.

Agresti, A. 1981. Measures of nominal–ordinal association. *J. Amer. Statist. Assoc.* **76**: 524–529.

Agresti, A. 1983a. A survey of strategies for modeling cross-classifications having ordinal variables. *J. Amer. Statist. Assoc.* **78**: 184–198.

Agresti, A. 1983b. Testing marginal homogeneity for ordinal categorical variables. *Biometrics* **39**: 505–510.

Agresti, A. 1983c. A simple diagonals-parameter symmetry and quasisymmetry model. *Statist. Probab. Lett.* **1**: 313–316.

Agresti, A. 1986. Applying R^2-type measures to ordered categorical data. *Technometrics* **28**: 133–138.

Agresti, A. 1988. A model for agreement between ratings on an ordinal scale. *Biometrics* **44**: 539–548.

Agresti, A. 1992a. Analysis of ordinal paired comparison data. *Appl. Statist.* **41**: 287–297.

Agresti, A. 1992b. A survey of exact inference for contingency tables. *Statist. Sci.* **7**: 131–153.

Agresti, A. 1993a. Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scand. J. Statist.* **20**: 63–71.

Agresti, A. 1993b. Distribution-free fitting of logit models with random effects of repeated categorical responses. *Statist. Med.* **12**: 1969–1987.

Agresti, A. 1995. Logit models and related quasi-symmetric loglinear models for comparing responses to similar items in a survey. *Sociol. Methods Res.* **24**: 68–95.

Agresti, A. 1999. Modelling ordered categorical data: Recent advances and future challenges. *Statist. Med.* **18**: 2191–2207.

Agresti, A. 2002. *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley.

Agresti, A., and C. Chuang. 1989. Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. *Comput. Statist. Data Anal.* **7**: 245–258.

Agresti, A., and B. A. Coull. 1996. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* **52**: 1103–1111.

Agresti, A., and B. A. Coull. 1998. Order-restricted inference for monotone trend alternatives in contingency tables. *Comput. Statist. Data Anal.* **28**: 139–155.

Agresti, A., and B. A. Coull. 2002. The analysis of contingency tables under inequality constraints. *J. Statist. Plann. Inference* **107**: 45–73.

Agresti, A., and A. Kezouh. 1983. Association models for multidimensional cross-classifications of ordinal variables. *Commun. Statist. Theory Methods* **12**: 1261–1276.

Agresti, A., and J. Lang. 1993a. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* **80**: 527–534.

Agresti, A., and J. Lang. 1993b. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* **49**: 131–139.

Agresti, A., and R. Natarajan. 2001. Modeling clustered ordered categorical data: A survey. *Internat. Statist. Rev.* **69**: 345–371.

Agresti, A., and D. Wackerly. 1977. Some exact conditional tests of independence for R × C cross-classification tables. *Psychometrika* **42**: 111–125.

Agresti, A., D. Wackerly, and J. Boyett. 1979. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* **44**: 75–84.

Agresti, A., C. Chuang, and A. Kezouh. 1987a. Order-restricted score parameters in association models for contingency tables. *J. Amer. Statist. Assoc.* **82**: 619–623.

Agresti, A., J. Schollenberger, and D. Wackerly. 1987b. Models for the probability of concordance in cross-classification tables. *Quality Quantity* **21**: 49–57.

Agresti, A., C. R. Mehta, and N. R. Patel. 1990. Exact inference for contingency tables with ordered categories. *J. Amer. Statist. Assoc.* **85**: 453–458.

Agresti, A., S. Lipsitz, and J. B. Lang. 1992. Comparing marginal distributions of large, sparse contingency tables. *Comput. Statist. Data Anal.* **14**: 55–73.

Agresti, A., M. Bini, B. Bertaccini, and E. Ryu. 2008. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* **64**: 1270–1275.

Aitchison, J., and S. D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**: 131–140.

Aït-Sidi-Allal, M. L., A. Baccini, and A. M. Mondot. 2004. A new algorithm for estimating the parameters and their asymptotic covariance in correlation and association models. *Comput. Statist. Data Anal.* **45**: 389–421.

Akritas, M. 1991. Limitations of the rank transform procedure: A study of repeated measures designs, Part I. *J. Amer. Statist. Assoc.* **86**: 457–460.

Akritas, M., and S. Arnold. 1994. Fully nonparametric hypotheses for factorial designs: I. Multivariate repeated measures designs. *J. Amer. Statist. Assoc.* **89**: 336–343.

Akritas, M. G., and E. Brunner. 1997. A unified approach to rank tests for mixed models. *J. Statist. Plann. Inference* **61**: 249–277.

Akritas, M. G., S. F. Arnold, and E. Brunner. 1997. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Amer. Statist. Assoc.* **92**: 258–265.

Albert, J. H. 1997. Bayesian testing and estimation of association in a two-way contingency table. *J. Amer. Statist. Assoc.* **92**: 685–693.

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**: 669–679.

Albert, J. H., and S. Chib. 2001. Sequential ordinal modeling with applications to survival data. *Biometrics* **57**: 829–836.

Albert, P. S. 1994. A Markov model for sequences of ordinal data from a relapsing-remitting disease. *Biometrics* **50**: 51–60.

Altham, P. M. E. 1969. Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's "exact" significance. test. *J. Roy. Statist. Soc.* **B 31**: 261–269.

Altham, P. M. E. 1975. Quasi-independent triangular contingency tables. *Biometrics* **31**: 233–238.

Ananth, C. V., and D. G. Kleinbaum. 1997. Regression models for ordinal responses: A review of methods and applications. *Internat. J. Epidemiol.* **26**: 1323–1333.

Andersen, E. B. 1973. Conditional inference for multiple-choice questionnaires. *Brit. J. Math. Statist. Psychol.* **26**: 31–44.

Andersen, E. B. 1980. *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

Andersen, E. B. 1995. Polytomous Rasch models and their estimation. In *Rasch Models: Foundations, Recent Developments, and Applications*, ed. G. Fischer and I. Molenaar. New York: Springer-Verlag, pp. 272–291.

Anderson, C. J. 1996. The analysis of three-way contingency tables by three-mode association models. *Psychometrika* **61**: 465–483.

Anderson, C. J. 2002. Analysis of multivariate frequency data by graphical models and generalizations of the multidimensional row–column association model. *Psychol. Methods* **7**: 446–467.

Anderson, C. J., and U. Böckenholt. 2000. Graphical regression models for polytomous variables. *Psychometrika* **65**: 497–509.

Anderson, C. J., and J. K. Vermunt. 2000. Log-multiplicative models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* **30**: 81–121.

Anderson, C. J., and H.-T. Yu. 2007. Log-multiplicative association models as item response models. *Psychometrika* **72**: 5–23.

Anderson, J. A. 1984. Regression and ordered categorical variables. *J. Roy. Statist. Soc.* **B 46**: 1–30.

Anderson, J. A., and P. R. Philips. 1981. Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.* **30**: 22–31.

Andrich, D. 1978. A rating formulation for ordered response categories. *Psychometrika* **43**: 561–573.

Andrich, D. 1979. A model for contingency tables having an ordered response classification. *Biometrics* **35**: 403–415.

Anscombe, F. J. 1981. *Computing in Statistical Science Through APL.* New York: Springer-Verlag.

Aranda-Ordaz, F. J. 1983. An extension of the proportional hazards model for grouped data. *Biometrics* **39**: 109–117.

Armitage, P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375–386.

Armstrong, B. G., and M. Sloan. 1989. Ordinal regression models for epidemiologic data. *Amer. J. Epidemiol.* **129**: 191–204.

Ashford, J. R. 1959. An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics* **15**: 573–581.

Bamber, D. 1975. The area under the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**: 387–415.

Banerjee, C., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.* **27**: 3–23.

Barnhart, H. X., and A. R. Sampson. 1994. Overview of multinomial models for ordinal data. *Commun. Statist. Theory Methods* **23**: 3395–3416.

Bartholomew, D. J. 1959. A test of homogeneity for ordered alternatives. *Biometrika* **46**: 36–48.

Bartholomew, D. J. 1980. Factor analysis for categorical data. *J. Roy. Statist. Soc.* **B 42**: 293–321.

Bartholomew, D. J. 1983. Latent variable models for ordered categorical data. *J. Econometrics* **22**: 229–243.

Bartolucci, F., and A. Farcomeni. 2009. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J. Amer. Statist. Assoc.* **104**: 816–831.

Bartolucci, F., and A. Forcina. 2002. Extended RC association models allowing for order restrictions and marginal modeling. *J. Amer. Statist. Assoc.* **97**: 1192–1199.

Bartolucci, F., and L. Scaccia. 2004. Testing for positive association in contingency tables with fixed margins. *Comput. Statist. Data Anal.* **47**: 195–210.

Bartolucci, F., A. Forcina, and V. Dardanoni. 2001. Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. *J. Amer. Statist. Assoc.* **96**: 1497–1505.

Bathke, A. 2005. Testing monotone effects of covariates in nonparametric mixed models. *J. Nonpar. Statist.* **17**: 423–439.

Bathke, A., and E. Brunner. 2003. A nonparametric alternative to analysis of covariance. In *Recent Advances and Trends in Nonparametric Statistics*, ed. M. G. Akritas and D. N. Politis. Boston: Elsevier, pp. 109–120.

Becker, M. 1989a. Models for the analysis of association in multivariate contingency tables. *J. Amer. Statist. Assoc.* **84**: 1014–1019.

Becker, M. 1989b. On the bivariate normal distribution and association models for ordinal categorical data. *Statist. Probab. Lett.* **8**: 435–440.

Becker, M. 1990a. Maximum likelihood estimation of the RC(M) association model. *Appl. Statist.* **39**: 152–167.

Becker, M. 1990b. Quasisymmetric models for the analysis of square contingency tables. *J. Roy. Statist. Soc.* **52**: 369–378.

Becker, M., and A. Agresti. 1992. Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statist. Med.* **11**: 101–114.

Becker, M., and C. C. Clogg. 1989. Analysis of sets of two-way contingency tables using association models. *J. Amer. Statist. Assoc.* **84**: 142–151.

Becker, W. E., and P. E. Kennedy. 1992. A graphical exposition of the ordered probit. *Econometric Theory* **8**: 127–131.

Beder, J. H., and R. C. Heim. 1990. On the use of ridit analysis. *Psychometrika* **55**: 603–616.

Bedrick, E. J., R. Christensen, and W. Johnson. 1996. A new perspective on priors for generalized linear models. *J. Amer. Statist. Assoc.* **91**: 1450–1460.

Beh, E. J. 1997. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical J.* **39**: 589–613.

Beh, E. J. 2001. Partitioning Pearson's chi-squared statistic for singly ordered two-way contingency tables. *Austral. New Zealand J. Statist.* **43**: 327–333.

Beh, E. J., and P. J. Davy. 2004. A non-iterative alternative to ordinal log-linear models. *J. Appl. Math. Decision Sci.* **8**: 67–86.

Beh, E. J., B. Simonetti, and L. DAmbra. 2007. Partitioning a non-symmetric measure of association for three-way contingency tables. *J. Multivariate Anal.* **98**: 1391–1411.

Bender, R., and A. Benner. 2000. Calculating ordinal regression models in SAS and S-Plus. *Biometrical J.* **42**: 677–699.

Bender, R., and U. Grouven. 1998. Using binary logistic regression models for ordinal data with non-proportional odds. *J. Clin. Epidemiol.* **51**: 809–816.

Bennett, S. 1983. Analysis of survival data by the proportional odds model. *Statist. Med.* **2**: 279–285.

Berger, V. W. 1998. Admissibility of exact conditional tests of stochastic order. *J. Statist. Plann. Inference* **66**: 39–50.

Berger, V. W., and H. Sackrowitz. 1997. Improving tests for superior treatment in contingency tables. *J. Amer. Statist. Assoc.* **92**: 700–705.

Berger, V. W., T. Permutt, and A. Ivanova. 1998. Convex hull test for ordered categorical data. *Biometrics* **54**: 1541–1550.

Berridge, D. M., and J. Whitehead. 1991. Analysis of failure time data with ordinal categories of response. *Statist. Med.* **10**: 1703–1710.

Best, N. G., D. J. Spiegelhalter, A. Thomas, and C. E. G. Brayne. 1996. Bayesian analysis of realistically complex models. *J. Roy. Statist. Soc.* A **159**: 323–342.

Bhapkar, V. P. 1968. On the analysis of contingency tables with a quantitative response. *Biometrics* **24**: 329–338.

Bhattacharya, B., and B. Nandram. 1996. Bayesian inference for multinomial populations under stochastic ordering. *J. Statist. Comput. Simul.* **54**: 145–163.

Birch, M. W. 1965. The detection of partial association. II: The general case. *J. Roy. Statist. Soc.* B **27**: 111–124.

Bishop, Y. M. M., and S. E. Fienberg. 1969. Incomplete two-dimensional contingency tables. *Biometrics* **25**: 119–128.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press (reprinted by Springer, New York: 2007).

Biswas, A., and K. Das. 2002. A Bayesian analysis of bivariate ordinal data: Wisconsin epidemiologic study of diabetic retinopathy revisited. *Statist. Med.* **21**: 549–559.

Bock, R. D. 1975. *Multivariate Statistical Methods in Behavioral Science.* McGraw-Hill.

Bock, R. D., and L. V. Jones. 1968. *The Measurement and Prediction of Judgement and Choice.* San Francisco: Holden-Day.

Böckenholt, U. 1999. Measuring change: Mixed Markov models for ordinal panel data. *Brit. J. Math. Statist. Psychol.* **52**: 125–136.

Böckenholt, U., and W. Dillon. 1997. Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika* **62**: 411–434.

Borgatta, E. F. 1968. My student, the purist: A lament. *Sociol. Q.* **9**: 29–34.

Borooah, V. K. 2002. *Logit and Probit: Ordered and Multinomial Models.* Thousand Oaks, CA: Sage.

Bradley, R. A., and M. E. Terry. 1952. Rank analysis of incomplete block designs. I: The method of paired comparisons. *Biometrika* **39**: 324–345.

Bradlow, E. T., and A. M. Zaslavsky. 1999. A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Amer. Statist. Assoc.* **94**: 43–52.

Brant, R. 1990. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**: 1171–1178.

Brockett, P. L., and A. Levine. 1977. On a characterization of ridits. *Ann. Statist.* **5**: 1245–1248.

Broemeling, L. D. 2009. *Bayesian Methods for Measures of Agreement*. London: Chapman & Hall.

Bross, I. D. J. 1958. How to use ridit analysis. *Biometrics* **14**: 18–38.

Brown, P. J., and P. W. K. Rundell. 1985. Kernel estimates for categorical data. *Technometrics* **27**: 293–299.

Brunk, H. D., W. E. Franck, D. L. Hanson, and R. V. Hogg. 1966. Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *J. Amer. Statist. Assoc.* **61**: 1067–1080.

Brunner, E., and F. Langer. 2000. Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical J.* **42**: 663–675.

Brunner, E., and U. Munzel. 2000. The nonparametric Behrens–Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical J.* **42**: 17–25.

Brunner, E., and M. L. Puri. 2001. Nonparametric methods in factorial designs. *Statist. Papers* **42**: 1–52.

Brunner, E., and M. L. Puri. 2002. A class of rank-score tests in factorial designs. *J. Statist. Plann. Inference* **103**: 331–360.

Brunner, E., M. L. Puri, and S. Sun. 1995. Nonparametric methods for stratified two-sample designs with application to multiclinic trials. *J. Amer. Statist. Assoc.* **90**: 1004–1014.

Brunner, E., H. Dette, and A. Munk. 1997. Box-type approximations in nonparametric factorial designs. *J. Amer. Statist. Assoc.* **92**: 1494–1502.

Brunner, E., U. Munzel, and M. L. Puri. 1999. Rank-score tests in factorial designs with repeated measures. *J. Multivariate Anal.* **70**: 286–317.

Brunner, E., S. Domhof, and F. Langer. 2002. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. Hoboken, NJ: Wiley.

Burridge, J. 1981. A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc.* B **43**: 41–45.

Carr, G. J., K. B. Hafner, and G. G. Koch. 1989. Analysis of rank measures of association for ordinal data from longitudinal studies. *J. Amer. Statist. Assoc.* **84**: 797–804.

Catalano, P. 1997. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statist. Med.* **16**: 883–900.

Chacko, V. J. 1966. Modified chi-square test for ordered alternatives. *Sankhya Ser. B* **28**, 185–190.

Chen, M.-H., and D. K. Dey. 2000. Bayesian analysis for correlated ordinal data analysis. In *Generalized Linear Models: A Bayesian Perspective*, ed. D. K. Dey, S. K. Ghosh, and B. K. Mallick. New York: Marcel Dekker, pp. 133–158.

Chib, S., and E. Greenberg. 1998. Analysis of multivariate probit models. *Biometrika* **85**: 347–361.

Chipman, H., and M. Hamada. 1996. Bayesian analysis of ordered categorical data from industrial experiments. *Technometrics* **38**: 1–10.

Choulakian, V. 1988. Exploratory analysis of contingency tables by loglinear formulation and generalizations of correspondence analysis. *Psychometrika* **53**: 235–250.

Chu, W., and Z. Ghahramani. 2005. Gaussian processes for ordinal regression. *J. Machine Learn. Res.* **6**: 1019–1041.

Chu, W., and S. S. Keerthi. 2007. Support vector ordinal regresion. *Neural Comput.* **19**: 792–815.

Chuang, C., and A. Agresti. 1986. A new model for ordinal pain data from a pharmaceutical study. *Statist. Med.* **5**: 15–20.

Chuang, C., D. Gheva, and C. Odoroff. 1985. Methods for diagnosing multiplicative-interaction models for two-way contingency tables. *Commun. Statist. Theory Methods* **14**: 2057–2080.

Chuang-Stein, C., and A. Agresti. 1997. A review of tests for detecting a monotone dose–response relationship with ordinal responses data. *Statist. Med.* **16**: 2599–2618.

Clayton, D. G. 1974. Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika* **61**: 525–531.

Clayton, D. G. 1976. An odds ratio comparison for ordered categorical data with censored observations. *Biometrika* **63**: 405–408.

Cliff, N., and J. A. Keats. 2002. *Ordinal Measurement in the Behavioral Sciences.* Mahwah, NJ: Lawrence Erlbaum.

Clogg, C. C. 1982a. Some models for the analysis of association in multiway cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **77**: 803–815.

Clogg, C. C. 1982b. Using association models in sociological research: Some examples. *Amer. J. Sociol.* **88**: 114–134.

Clogg C. C., and E. S. Shihadeh. 1994. *Statistical Models for Ordinal Variables.* Thousand Oaks, CA: Sage.

Cochran, W. G. 1954. Some methods of strengthening the common χ^2 tests. *Biometrics* **10**: 417–451.

Cochran, W. G. 1955. A test of a linear function of the deviations between observed and expected numbers. *J. Amer. Statist. Assoc.* **50**: 377–397.

Cochran, W. G. 1968. Effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**: 295–313.

Cohen, A., and H. B. Sackrowitz. 1991. Tests for independence in contingency tables with ordered alternatives. *J. Multivariate Anal.* **36**: 56–67.

Cohen, A., and H. B. Sackrowitz. 1992. An evaluation of some tests of trend in contingency tables. *J. Amer. Statist. Assoc.* **87**: 470–475.

Cohen, A., and H. B. Sackrowitz. 1998. Directional tests for one-sided alternatives in multivariate models. *Ann. Statist.* **26**: 2321–2378.

Cohen, A., H. B. Sackrowitz, and M. Sackrowitz. 2000. Testing whether treatment is "better" than control with ordered categorical data: An evaluation of new methodology. *Statist. Med.* **19**: 2699–2712.

Cohen, A., D. Madigan, and H. B. Sackrowitz. 2003. Effective directed tests for models with ordered categorical data. *Austral. New Zealand J. Statist.* **45**: 285–300.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**: 37–46.

Cole, S. R., and C. V. Ananth. 2001. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *Internat. J. Epidemiol.* **30**: 1379–1382.

Cole, S. R., P. D. Allison, and C. V. Ananth. 2004. Estimation of cumulative odds ratios. *Ann. Epidemiol.* **14**: 172–178.

Colombi, R., and A. Forcina. 2001. Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika* **88**: 1007–1019.

Congdon, P. 2005. *Bayesian Models for Categorical Data*. Hoboken, NJ: Wiley.

Conover, W. J., and R. L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *Amer. Statist.* **35**: 124–129.

Corcoran, C., C. Mehta, and P. Senchaudhuri. 2000. Power comparisons for tests of trend in dose–response studies. *Statist. Med.* **19**: 3037–3050.

Coste, J., E. Walter, D. Wasserman, and A. Venot. 1997. Optimal discriminant analysis for ordinal responses. *Statist. Med.* **16**: 561–569.

Coull, B. A., and A. Agresti. 2000. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56**: 73–80.

Coull, B. A., and A. Agresti. 2003. A class of generalized log-linear models with random effects. *Statist. Model.* **3**: 251–271.

Cowles, M. K. 1996. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statist. Comput.* **6**: 101–111.

Cowles, M. K., B. P. Carlin, and J. E. Connett. 1996. Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *J. Amer. Statist. Assoc.* **91**: 86–98.

Cox, C. 1988. Multinomial regression models based on continuation ratios. *Statist. Med.* **7**: 435–441.

Cox, C. 1995. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statist. Med.* **14**: 1191–1203.

Cox, C., and C. Chuang. 1984. A comparison of chi-square partitioning and two logit analyses of ordinal pain data from a pharaceutical study. *Statist. Med.* **3**: 273–285.

Croon, M. A. 1990. Latent class analysis with ordered classes. *Brit. J. Math. Statist. Psychol.* **43**: 171–192.

Crouchley, R. 1995. A random-effects model for ordered categorical data. *J. Amer. Statist. Assoc.* **90**: 489–498.

Dale, J. R. 1984. Local versus global association for bivariate ordered responses. *Biometrika* **71**: 507–514.

Dale, J. R. 1986. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**: 909–917.

Daniels, H. E. 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33**: 129–135.

Dardanoni, V., and A. Forcina. 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Amer. Statist. Assoc.* **93**: 1112–1123.

Davis, C. S. 1988. Estimation of row and column scores in the linear-by-linear association model for two-way ordinal contingency tables. *Proc. 13th Annual SAS Users Group International*. Cary, NC: SAS Institute.

Davis, J. A. 1967. A partial coefficient for Goodman and Kruskal's gamma. *J. Amer. Statist. Assoc.* **62**: 189–193.

de Falguerolles, A., S. Jmel, and J. Whittaker. 1995. Correspondence analysis and association models constrained by a conditional independence graph. *Psychometrika* **60**: 161–180.

de Rooij, M. 2001. Distance association models for the analysis of repeated transition frequency tables. *Statist. Neerlandica* **55**: 157–181.

de Rooij, M., and W. J. Heiser. 2005. Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika* **70**: 99–122.

DeSantis, S. M., E. A. Houseman, B. A. Coull, A. Stemmer-Rachamimov, and R. Betensky. 2008. A penalized latent class model for ordinal data. *Biostatistics* **9**: 249–262.

DiPrete, T. A. 1990. Adding covariates to loglinear models for the study of social mobility. *Amer. Sociol. Rev.* **55**: 757–773.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser. 2004. A log-linear approach for modelling ordinal paired comparison data on motives to start a PhD programme. *Statist. Model.* **4**: 181–193.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser. 2007. A paired comparison approach for the analysis of sets of Likert-scale responses *Statist. Model.* **7**: 3–28.

Dong, J., and J. S. Simonoff. 1995. A geometric combination estimator for *d*-dimensional ordinal sparse contingency tables. *Ann. Statist.* **23**: 1143–1159.

Dos Santos, D. M., and D. M. Berridge. 2000. A continuation ratio random effects model for repeated ordinal responses. *Statist. Med.* **19**: 3377–3388.

Douglas, R., and S. E. Fienberg. 1990. An overview of dependency models for cross-classified categorical data involving ordinal variables. In *Topics in Statistical Dependence*, ed. H. W. Block, A. R. Sampson, and T. H. Savits. Hayward, CA: Institute of Mathematical Statistics.

Douglas, R., S. E. Fienberg, M.-L. T. Lee, A. R. Sampson, and L. R. Whitaker. 1990. Positive dependence concepts for ordinal contingency tables. *Topics in Statistical Dependence*, ed. H. W. Block, A. R. Sampson, and T. H. Savits. Hayward, CA: Institute of Mathematical Statistics, pp. 189–202.

Drasgow, F. 1988. Polychoric and polyserial correlations. *Ency. Statist. Sci.* **7**: 69–74.

Duncan, O. D. 1979. How destination depends on origin in the occupational mobility table. *Amer. J. Sociol.* **84**: 793–803.

Duncan, O. D., and J. A. McRae, Jr. 1979. Multiway contingency analysis with a scaled response or factor. *Sociol. Methodol.* **10**: 66–85.

Dykstra, R. L., and C. C. Lee. 1991. Multinomial estimation procedures for isotonic cones. *Statist. Probab. Lett.* **11**: 155–160.

Dykstra, R. L., and J. Lemke. 1988. Duality of *I* projections and maximum likelihood estimation for log-linear models under cone constraints. *J. Amer. Statist. Assoc.* **83**: 546–554.

Dykstra, R. L., S. Kochar, and T. Robertson. 1995. Inference for likelihood ratio ordering in the two-sample problem. *J. Amer. Statist. Assoc.* **90**: 1034–1040.

Edwardes, M. D. deB. 1993. Kendall's τ is equal to the correlation coefficient for the BVE distribution. *Statist. Probab. Lett.* **17**: 415–419.

Edwardes, M. D. deB. 1997. Univariate random cut-points theory for the analysis of ordered categorical data. *J. Amer. Statist. Assoc.* **92**: 1114–1123.

Edwardes, M. D. deB. 2002. Distribution-free tests for cluster samples of ordinal responses. *J. Statist. Plann. Inference* **105**: 393–404.

Ekholm, A., J. Jokinen, J. W. McDonald, and P. W. F. Smith. 2003. Joint regression and association modeling of longitudinal ordinal data. *Biometrics* **59**: 795–803.

Ekström, J. 2009. *Contributions to the Theory of Measures of Association for Ordinal Variables*. Unpublished Ph.D. dissertation, Uppsala Universitet, Sweden.

El Barmi, H., and R. L. Dykstra. 1995. Testing for and against a set of linear inequality constraints in a multinomial setting. *Canad. J. Statist.* **23**: 131–143.

Epstein, L. D., and S. E. Fienberg. 1992. Bayesian estimation in multidimensional contingency tables. In *Bayesian Analysis in Statistics and Econometrics*. New York: Springer-Verlag, pp. 27–41.

Etzioni, R. D., S. E. Fienberg, Z. Gilula, and S. J. Haberman. 1994. Statistical models for the analysis of ordered categorical data in public health and medical research. *Statist. Methods Med. Res.* **3**: 179–204.

Evans, M., Z. Gilula, and I. Guttman. 1993. Computational issues in the Bayesian analysis of categorical data: Log-linear and Goodman's RC model. *Statist. Sinica* **3**: 391–406.

Evans, M., Z. Gilula, I. Guttman, and T. Swartz. 1997. Bayesian analysis of stochastically ordered distributions of categorical variables. *J. Amer. Statist. Assoc.* **92**: 208–214.

Ezzet, F., and J. Whitehead. 1991. A random effects model for ordinal responses from a crossover trial. *Statist. Med.* **10**: 901–907.

Fagot, R. F. 1994. An ordinal coefficient of relational agreement for multiple judges. *Psychometrika* **59**: 241–251.

Fahrmeir, L., and G. Tutz. 1994. Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Amer. Statist. Assoc.* **89**: 1438–1449.

Fahrmeir, L., and G. Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. New York: Springer-Verlag.

Farewell, V. T. 1982. A note on regression analysis of ordinal data with variability of classification. *Biometrika* **69**: 533–538.

Fay, M. P., and C. Gennings. 1996. Non-parametric two-sample tests for repeated ordinal responses. *Statist. Med.* **15**: 429–442.

Feldmann, U., and I. Steudel. 2000. Methods of ordinal classification applied to medical scoring systems. *Statist. Med.* **19**: 575–586.

Fielding, A. 1999. Why use arbitrary points scores?: Ordered categories in models of educational progress. *J. Roy. Statist. Soc.* A **162**: 303–328.

Fielding, A., and M. Yang. 2005. Generalized linear mixed models for ordered responses in complex multilevel structures: Effects beneath the school or college in education. *J. Roy. Statist. Soc.* A **168**: 159–183.

Fielding, A., M. Yang, and H. Goldstein. 2003. Multilevel ordinal models for examination grades. *Statist. Model.* **3**: 127–153.

Fienberg, S. E. 1980. *The Analysis of Cross-Classified Categorical Data*, 2nd ed. Cambridge, MA: MIT Press (reprinted by Springer, New York, 2007).

Fienberg, S. E., and P. W. Holland. 1973. Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68**: 683–690.

Fienberg, S. E., and W. M. Mason. 1979. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociol. Methodol.* **10**: 1–67.

Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **33**: 613–619.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72**: 323–327.

Francom, S. F., C. Chuang-Stein, and J. R. Landis. 1989. A log-linear model for ordinal data to characterize differential change among treatments. *Statist. Med.* **8**: 571–582.

Frank, E., and M. Hall. 2001. A simple approach to ordinal classification. In *Machine Learning*. Springer Lecture Notes in Computer Science. New York: Springer-Verlag, pp. 145–156.

Freedman, D. A. 1963. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**: 1386–1403.

Freidlin, B., M. J. Podgor, and J. L. Gastwirth. 1999. Efficiency robust tests for survival or ordered categorical data. *Biometrics* **55**: 883–886.

Fu, L., and D. G. Simpson. 2002. Conditional risk models for ordinal response data: Simultaneous logistic regression analysis and generalized score tests. *J. Statist. Plann. Inference* **108**: 201–217.

Galindo-Garre, F., and J. K. Vermunt. 2004. The order-restricted association model: Two estimation algorithms and issues in testing. *Psychometrika* **69**: 641–654.

Gans, L. P., and C. A. Robertson. 1981. Distributions of Goodman and Kruskal's gamma and Spearman's rho in 2×2 tables for small and moderate sample sizes. *J. Amer. Statist. Assoc.* **76**: 942–946.

Gao, W., and S. Kuriki. 2006. Testing marginal homogeneity against stochastically ordered marginals for $r \times r$ contingency tables. *J. Multivariate Anal.* **97**: 1330–1341.

Gautam, S. 1997. Test for linear trend in $2 \times K$ table with open-ended categories. *Biometrics* **53**: 1163–1169.

Gautam, S., A. R. Sampson, and H. Singh. 2001. Iso-chi-squared testing of $2 \times K$ ordered tables. *Canad. J. Statist.* **29**: 609–619.

Gelfand, A. E., and L. Kuo. 1991. Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**: 657–666.

Gelfand, A. E., A. F. M. Smith, and T.-M. Lee. 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87**: 523–532.

Genter, F. C., and V. T. Farewell. 1985. Goodness-of-link testing in ordinal regression models. *Canad. J. Statist.* **13**: 37–44.

Gill, J., and G. Casella. 2009. Nonparametric priors for ordinal Bayesian social science models: Specification and estimation. *J. Amer. Statist. Assoc.* **104**: 453–464.

Gilula, Z. 1982. A note on the analysis of association in cross classifications having ordered categories. *Commun. Statist. Theory Methods* **11**: 1233–1240.

Gilula, Z. 1984. On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika* **71**: 523–529.

Gilula, Z. 1986. Grouping and association in two-way contingency tables: A canonical correlation analytic approach. *J. Amer. Statist. Assoc.* **81**: 773–779.

Gilula, Z., and S. Haberman. 1986. Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* **81**: 780–788.

Gilula, Z., and S. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.* **83**: 760–771.

Gilula, Z., and Y. Ritov. 1990. Inferential ordinal correspondence analysis: Motivation, derivation and limitations. *Internat. Statist. Rev.* **58**: 99–108.

Gilula, Z., A. Krieger, and Y. Ritov. 1988. Ordinal association in contingency tables: Some interpretive aspects. *J. Amer. Statist. Assoc.* **83**: 540–545.

Glewwe, P. 1997. A test of the normality assumption in ordered probit model. *Econometric Rev.* **16**: 1–19.

Glonek, G. 1996. A class of regression models for multivariate categorical responses. *Biometrika* **83**: 15–28.

Glonek, G. F. V., and P. McCullagh. 1995. Multivariate logistic models. *J. Roy. Statist. Soc.* **B 57**: 533–546.

Gonin, R., S. R. Lipsitz, G. M. Fitzmaurice, and G. Molenberghs. 2000. Regression modelling of weighted κ by using generalized estimating equations. *J. Roy. Statist. Soc.* **C 49**: 1–18.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Goodman, L. A. 1968. The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Assoc.* **63**: 1091–1131.

Goodman, L. A. 1972. Some multiplicative models for the analysis of cross-classified data. In *Proc. 6th Berkeley Symposium*, ed. L. Le Cam et al., vol. **1**. Berkeley, CA: University of California Press, pp. 649–696.

Goodman, L. A. 1979a. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74**: 537–552.

Goodman, L. A. 1979b. Multiplicative models for square contingency tables with ordered categories. *Biometrika* **66**: 413–418.

Goodman, L. A. 1979c. Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables. *Amer. J. Sociol.* **84**: 804–819.

Goodman, L. A. 1981a. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76**: 320–334.

Goodman, L. A. 1981b. Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68**: 347–355.

Goodman, L.A. 1981c. Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. *Sociol. Methodol.* **12**: 193–239.

Goodman, L. A. 1983. The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* **39**: 149–160.

Goodman, L. A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13**: 10–69.

Goodman, L. A. 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Internat. Statist. Rev.* **54**: 243–309.

Goodman, L. A. 1991. Measures, models, and graphical displays in the analysis of cross-classified data. *J. Amer. Statist. Assoc.* **86**: 1085–1111.

Goodman, L. A. 1996. A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *J. Amer. Statist. Assoc.* **91**: 408–427.

Goodman, L. A. 2000. The analysis of cross-classified data: Notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future. In *Statistics for the 21st Century*, ed. C. R. Rao and G. J. Székely. New York: Marcel Dekker, pp. 189–231.

Goodman, L. A., and W. H. Kruskal. 1954. Measures of association for cross classifications. *J. Amer. Statist. Assoc.* **49**: 732–764.

Goodman, L. A., and W. H. Kruskal. 1963. Measures of association for cross classifications. III: Approximate sampling theory. *J. Amer. Statist. Assoc.* **58**: 310–364.

Goodman, L. A., and W. H. Kruskal. 1972. Measures of association for cross classifications. IV: Simplification of asymptotic variances. *J. Amer. Statist. Assoc.* **67**: 415–421.

Graham, P., and R. Jackson. 1993. The analysis of ordinal agreement data: Beyond weighted kappa. *J. Clin. Epidemiol.* **46**: 1055–1062.

Graubard, B. I., and E. L. Korn. 1987. Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**: 471–476.

Greenacre, M. J. 2007. *Correspondence Analysis in Practice*, 2nd ed. London: Chapman & Hall.

Greenland, S. 1994. Alternative models for ordinal logistic regression. *Statist. Med.* **13**: 1665–1677.

Greenwood, C., and V. T. Farewell. 1988. A comparison of regression models for ordinal data in an analysis of transplanted-kidney function. *Canad. J. Statist.* **16**: 325–335.

Grilli, L. 2005. The random-effects proportional hazards models with grouped survival data: A comparison between the grouped continuous and continuation ratio varsions. *J. Roy. Statist. Soc.* **A 168**: 83–94.

Grilli, L., and C. Rampichini. 2002. Specification issues in stratified variance component ordinal response models. *Statist. Model.* **2**: 251–264.

Grilli, L., and C. Rampichini. 2003. Alternative specifications of multivariate multilevel probit ordinal response models. *J. Educ. Behav. Statist.* **28**: 31–44.

Grilli, L., and C. Rampichini. 2007. Multilevel factor models for ordinal variables. *Struct. Equation Model.* **14**: 1–25.

Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* **25**: 489–504.

Gross, S. T. 1981. On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *J. Amer. Statist. Assoc.* **76**: 935–941.

Grove, D. M. 1980. A test of independence against a class of ordered alternatives in a $2 \times c$ contingency table. *J. Amer. Statist. Assoc.* **75**: 454–459.

Grove, D. M. 1984. Positive association in a two-way contingency table: Likelihood ratio tests. *Commun. Statist. Theory Methods* **13**: 931–945.

Grove, D. M. 1986. Positive association in a two-way contingency table: A numerical study. *Commun. Statist. Simul. Comput.* **15**: 633–648.

Guisan, A., and F. E. Harrell. 2000. Ordinal response regression models in ecology. *J. Vegetation Sci.* **11**: 617–626.

Gurland, J., I. Lee, and P. A. Dahm. 1960. Polychotomous quantal response in biological assay. *Biometrics* **16**: 382–398.

Haber, M. 1985. Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.* **3**: 1–10.

Haberman, S. J. 1974. Log-linear models for frequency tables with ordered classifications. *Biometrics* **36**: 589–600.

Haberman, S. J. 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Ann. Statist.* **9**: 1178–1186.

Haberman, S. J. 1995. Computation of maximum likelihood estimates in association models. *J. Amer. Statist. Assoc.* **90**: 1438–1446.

Halperin, M., M. I. Hamdy, and P. F. Thall. 1989. Distribution-free confidence intervals for a parameter of Wilcoxon–Mann–Whitney type for ordered categories and progressive censoring. *Biometrics* **45**: 509–521.

Hamada, M., and C. F. J. Wu. 1990. A critical look at accumulation analysis and related methods. *Technometrics* **32**: 119–130.

Hartzel, J., I.-M. Liu, and A. Agresti. 2001a. Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Comput. Statist. Data Anal.* **35**: 429–449.

Hartzel, J., A. Agresti, and B. Caffo. 2001b. Multinomial logit random effects models. *Statist. Model.* **1**: 81–102.

Harville, D., and R. W. Mee. 1984. A mixed-model procedure for analyzing ordered categorical data. *Biometrics* **40**: 393–408.

Hastie, T., and R. Tibshirani. 1987. Non-parametric logistic and proportional odds regression. *Appl. Statist.* **36**: 260–276.

Hastie, T. J., J. L. Botha, and C. M. Schnitzler. 1989. Regression with an ordered categorical response. *Statist. Med.* **8**: 785–794.

Hausman, J., A. Lo, and A. C. MacKinlay. 1992. An ordered probit analysis of stock transaction prices. *J. Financial Econ.* **31**: 319–379.

Hawkes, R. K. 1971. The multivariate analysis of ordinal measures. *Amer. J. Sociol.* **76**: 908–926.

Heagerty, P. J., and S. L. Zeger. 1996. Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* **91**: 1024–1036.

Heagerty, P. J., and S. L. Zeger. 2000a. Multivariate continuation ratio models: Connections and caveats. *Biometrics* **56**: 719–732.

Heagerty, P. J., and S. L. Zeger. 2000b. Marginalized multilevel models and likelihood inference. *Statist. Sci.* **15**: 1–26.

Hedeker, D., and R. D. Gibbons. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**: 933–944.

Hedeker, D., and R. D. Gibbons. 2006. *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.

Hedeker, D., and R. J. Mermelstein. 1998. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behav. Res.* **33**: 427–455.

Hedeker, D., O. Siddiqui, and F. B. Hu. 2000. Random-effects analysis of correlated grouped-time survival data. *Statist. Methods Med. Res.* **9**: 161–179.

Hedeker, D., M. Berbaum, and R. J. Mermelstein. 2006. Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *J. Probab. Statist. Sci.* **4**: 1–20.

Hemker, B. T., L. A. van der Ark, and K. Sijtsma. 2001. On measurement properties of continuation ratio models. *Psychometrika* **66**: 487–506.

Herbrich, R., T. Graepel, and K. Obermayer. 1999. Support vector learning for ordinal regression. *Artificial Neural Networks* **1**: 97–102.

Hildebrand, D. K., J. D. Laing, and H. Rosenthal. 1977. *Analysis of Ordinal Data.* Thousand Oaks, CA: Sage.

Hilton, J. F. 1996. The appropriateness of the Wilcoxon test in ordinal data. *Statist. Med.* **15**: 631–645.

Hilton, J. F., and C. R. Mehta. 1993. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* **49**: 609–616.

Hilton, J. F., C. R. Mehta, and N. R. Patel. 1994. An algorithm for conducting Smirnov tests. *Comput. Statist. Data Anal.* **17**: 351–361.

Hirji, K. F. 1992. Computing exact distributions for polytomous response data. *J. Amer. Statist. Assoc.* **87**: 487–492.

Hirji, K. F. 2005. *Exact Analysis of Discrete Data.* London: Chapman & Hall.

Hirotsu, C. 1982. Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* **69**: 567–577.

Hirotsu, C. 1983. Defining the pattern of association in two-way contingency tables. *Biometrika* **70**: 579–589.

Hobert, J. P., and G. Casella. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**: 1461–1473.

Hollander, M., and D. A. Wolfe. 1999. *Nonparametric Statistical Methods*, 2nd ed. New York: Wiley.

Holtbrügge, W., and M. Schumacher. 1991. A comparison of regression models for the analysis of ordered categorical data. *Appl. Statist.* **40**: 249–259.

Horvath, T., and P. Vojtas. 2006. Ordinal classification with monotonicity constraints. In *Advances in Data Mining.* Springer Lecture Notes in Computer Science. New York: Springer-Verlag, pp. 217–225.

Hout, M., O. D. Duncan, and M. E. Sobel. 1987. Association and heterogeneity: Structural models of similarities and differences. *Sociol. Methodol.* **17**: 145–184.

Huang, G.-H., K. Bandeen-Roche, and G. S. Rubin. 2002. Building marginal models for multiple ordinal measurements. *J. Roy. Statist. Soc.* **C 51**: 37–57.

Iliopoulos, G., M. Kateri, and I. Ntzoufras. 2007. Bayesian estimation of unrestricted and order-restricted association models for a two-way contingency table. *Comput. Statist. Data Anal.* **51**: 4643–4655.

Iliopoulos, G., M. Kateri, and I. Ntzoufras. 2009. Bayesian model comparison for the order restricted RC association model. *Psychometrika* **74**: 561–587.

Ishii-Kuntz, M. 1994. *Ordinal Log-Linear Models.* Thousand Oaks, CA: Sage.

Ishwaran, H. 2000. Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *Canad. J. Statist.* **28**: 715–730.

Ishwaran, H., and C. Gatsonis. 2000. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canad. J. Statist.* **28**: 731–750.

Ivanova, A., and V. W. Berger. 2001. Drawbacks to integer scoring for ordered categorical data. *Biometrics* **57**: 567–570.

Jansen, J. 1990. On the statistical analysis of ordinal data when extravariation is present. *Appl. Statist.* **39**: 74–85.

Jansen, M. E. 1984. Ridit analysis: A review. *Statist. Neerlandica* **38**: 141–158.

Jewell, N. P., and J. D. Kalbfleisch. 2004. Maximum likelihood estimation of ordered multi-nomial parameters. *Biostatistics* **5**: 291–306.

Joffe, M. M., and S. Greenland. 1995. Standardized estimates from categorical regression models. *Statist. Med.* **14**: 2131–2141.

Johnson, T. R. 2007. Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika* **72**: 489–504.

Johnson, V. E. 1996. On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *J. Amer. Statist. Assoc.* **91**: 42–51.

Johnson, V. E., and J. H. Albert 1999. *Ordinal Data Modeling*. New York: Springer.

Jokinen, J., J. W. McDonald, and P. W. F. Smith. 2006. Meaningful regression and association models for clustered ordinal data. *Sociol. Methodol.* **36**: 173–199.

Jöreskog, K. G. 1994. On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika* **59**: 381–389.

Jung, S.-H., and Kang, S.-H. 2001. Tests for $2 \times K$ contingency tables with clustered ordered categorical data. *Statist. Med.* **20**: 785–794.

Kaciroti, N. A., T. E. Raghunathan, M. A. Schork, N. M. Clark, and M. Gong. 2006. A Bayesian approach for clustered longitudinal ordinal outcome with nonignorable missing data: Evaluation of an asthma education program. *J. Amer. Statist. Assoc.* **101**: 435–446.

Kateri, M., and A. Agresti. 2007. A class of ordinal quasi-symmetry models for square contingency tables. *Comput. Statist. Data Anal.* **77**: 598–603.

Kateri, M., and G. Iliopoulos. 2003. On collapsing categories in two-way contingency tables. *Statistics*: **37**: 443–455.

Kateri, M., R. Ahmad, and T. Papaioannou. 1998. New features in the class of association models. *Appl. Stoch. Models Data Anal.* **14**: 125–136.

Kateri, M., A. Nicolaou, and I. Ntzoufras. 2005. Bayesian inference for the RC(m) association model. *J. Comput. Graph. Statist.* **23**: 116–138.

Kauermann, G. 2000. Modeling longitudinal data with ordinal response by varying coefficients. *Biometrics* **56**: 692–698.

Kauermann, G., and G. Tutz. 2003. Semi- and nonparametric modeling of ordinal data. *J. Comput. Graph. Statist.* **12**: 176–196.

Kawaguchi, A., and G. G. Koch. 2010. Multivariate Mann-Whitney estimators for the comparison of two treatments in a three period crossover study with randomly missing data. *J. Biopharm. Statist.* **21**: to appear.

Kaufmann, H. 1988. On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models. *Metrika* **35**: 291–313.

Kedem, B., and K. Fokianos. 2002. *Regression Models for Time Series Analysis*. Hoboken, NJ: Wiley.

Keen, A., and B. Engel. 1997. Analysis of a mixed model for ordinal data by iterative reweighted REML. *Statist. Neerlandica* **51**: 129–144.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* **30**: 81–93.

Kendall, M. G. 1945. The treatment of ties in rank problems. *Biometrika* **33**: 239–251.

Kendall, M. G. 1970. *Rank Correlation Methods*, 4th ed. London: Charles Griffin.

Kendall, M., and A. Stuart. 1979. *The Advanced Theory of Statistics*, vol. **2**: *Inference and Relationship*, 4th ed. New York: Macmillan.

Kenward, M. G., and B. Jones. 1991. The analysis of categorical data from cross-over trials using a latent variable model. *Statist. Med.* **10**: 1607–1619.

Kenward, M. G., E. Lesaffre, and G. Molenberghs. 1994. An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**: 945–953.

Kim, D., and A. Agresti. 1997. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Comput. Statist. Data Anal.* **24**: 89–104.

Kim, J. 1975. Multivariate analysis of ordinal variables. *Amer. J. Sociol.* **81**: 261–298.

Kim, J.-H. 2003. Assessing practical significance of the proportional odds assumption. *Statist. Probab. Lett.* **65**: 233–239.

Kim, K. 1995. A bivariate cumulative probit regression model for ordered categorical data. *Statist. Med.* **14**: 1341–1352.

Kimeldorf, G., and A. R. Sampson. 1989. A framework for positive dependence. *Ann. Inst. Statist. Math.* **41**: 31–45.

Kimeldorf, G., A. R. Sampson, and L. R. Whitaker. 1992. Min and max scorings for two-sample ordinal data. *J. Amer. Statist. Assoc.* **87**: 241–247.

Klingenberg, B., A. Solari, L. Salmaso, and F. Pesarin. 2009. Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics* **65**: 452–462.

Klotz, J. 1966. The Wilcoxon, ties, and the computer. *J. Amer. Statist. Assoc.* **61**: 772–787.

Klotz, J. 1980. A modified Cochran–Friedman test with missing observations and ordered categorical data. *Biometrics* **36**: 665–670.

Klotz, J., and J. Teng. 1977. One-way layout for counts and the exact enumeration of the Kruskal–Wallis *H* distribution with ties. *J. Amer. Statist. Assoc.* **72**: 165–169.

Koch, G. G., and D. W. Reinfurt. 1971. The analysis of categorical data from mixed models. *Biometrics* **27**: 157–173.

Koch, G. G., J. R. Landis, J. L. Freeman, D. H. Freeman, and R. G. Lehnen. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics.* **33**: 133–158.

Koch, G. G., I. A. Amara, G. W. Davis, and D. B. Gillings. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* **38**: 563–595.

Koch, G. G., C. M. Tangen, J.-W. Jung, and I. A. Amara. 1998. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statist. Med.* **17**: 1863–1892.

Kolassa, J. 1995. A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Statist. Med.* **14**: 1577–1581.

Kosorok, M. R., and W.-H. Chao. 1996. The analysis of longitudinal ordinal response data in continuous time. *J. Amer. Statist. Assoc.* **91**: 807–817.

Kottas, A., P. Müller, and F. Quintana. 2005. Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.* **14**: 610–625.

Kruskal, W. H. 1952. A nonparametric test for the several sample problem. *Ann. Math. Statist.* **23**: 525–540.

Kruskal, W. H. 1957. Historical notes on the Wilcoxon unpaired two-sample test. *J. Amer. Statist. Assoc.* **52**: 356–360.

Kruskal, W. H. 1958. Ordinal measures of association. *J. Amer. Statist. Assoc.* **53**: 814–861.

Kruskal, W. H., and W. A. Wallis. 1952. The use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* **47**: 583–621.

Kuss, O. 2006. On the estimation of the stereotype regression model. *Comput. Statist. Data Anal.* **50**: 1877–1890.

Kvist, T., H. Gislason, and P. Thyregod. 2000. Using continuation-ratio logits to analyze the variation of the age composition of fish catches. *J. Appl. Statist.* **27**: 303–319.

Läärä, E., and J. N. S. Matthews. 1985. The equivalence of two models for ordinal data. *Biometrika* **72**: 206–207.

Labovitz, S. 1970. The assignment of numbers to rank order categories. *Amer. Sociol. Rev.* **35**: 515–524.

Lall, R., M. J. Campbell, S. J. Walters, and K. Morgan. 2002. A review of ordinal regression models applied on health-related quality of life assessments. *Statist. Methods Med. Res.* **11**: 49–67.

Lancaster, H. O. 1949. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika* **36**: 117–129.

Lancaster, H. O. 1969. *The Chi-Squared Distribution.* New York: Wiley.

Lancaster, H. O., and M. A. Hamdan. 1964. Estimation of the correlation coefficient in contingency tables with possible nonmetrical characters. *Psychometrika* **29**: 383–391.

Landis, J. R., and G. G. Koch. 1977a. The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–174.

Landis, J. R., and G. G. Koch. 1977b. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**: 363–374.

Landis, J. R., E. R. Heyman, and G. G. Koch. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Internat. Statist. Rev.* **46**: 237–254.

Landis, J. R., M. E. Miller, C. S. Davis, and G. G. Koch. 1988. Some general methods for the analysis of categorical data in longitudinal studies. *Statist. Med.* **7**: 109–137.

Lang, J. B. 1996. Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.* **24**: 726–752.

Lang, J. B. 1999. Bayesian ordinal and binary regression models with a parametric family of mixture links. *Comput. Statist. Data Anal.* **31**: 59–87.

Lang, J. B. 2004. Multinomial-Poisson homogeneous models for contingency tables. *Ann. Statist.* **32**: 340–383.

Lang, J. B. 2005. Homogeneous linear predictor models for contingency tables. *J. Amer. Statist. Assoc.* **100**: 121–134.

Lang, J. B. 2008. Score and profile likelihood confidence intervals for contingency table parameters. *Statist. Med.* **27**: 5975–5990.

Lang, J. B., and A. Agresti. 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**: 625–632.

Lang, J. B., and S. R. Eliason. 1997. The application of association-marginal models to the study of social mobility. *Sociol. Methods Res.* **26**: 183–213.

Lang, J. B., J. W. McDonald, and P. W. F. Smith. 1999. Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *J. Amer. Statist. Assoc.* **94**: 1161–1171.

Lapp, K., G. Molenberghs, and E. Lesaffre. 1998. Models for the association between ordinal variables. *Comput. Statist. Data Anal.* **28**: 387–411.

Larichev, O. I., and H. M. Moshkovich. 1994. An approach to ordinal classification problems. *Internat. Trans. Operations Res.* **1**: 375–385.

Lauritzen, S. L., and N. Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**: 31–57.

Lawrence, E., D. Bingham, C. Liu, and V. N. Nair. 2008. Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics* **50**: 182–191.

Lee, C. C. 1987. Chi-squared tests for and against an order restriction on multinomial parameters. *J. Amer. Statist. Assoc.* **82**: 611–618.

Lee, K., and M. J. Daniels. 2007. A class of Markov models for longitudinal ordinal data. *Biometrics* **63**: 1060–1067.

Lee, M.-K., H.-H. Song, S.-H. Kang, and C. W. Ahn. 2002. The determination of sample sizes in the comparison of two multinomial proportions from ordered categories. *Biometrical J.* **44**: 395–409.

Lee, S.-Y., W.-Y. Poon, and P. M. Bentler. 1992. Structural equation models with continuous and polytomous variables. *Psychometrika* **57**: 89–105.

Lehmann, E. L. 1966. Some concepts of dependence. *Ann. Math. Statist.* **37**: 1137–1153.

Lehmann, E. L. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Leonard, T. 1973. A Bayesian method for histograms. *Biometrika* **60**: 297–308.

Lesaffre, E., G. Molenberghs, and L. Dewulf. 1998. Effect of dropouts in a longitudinal study: An application of a repeated ordinal model. *Statist. Med.* **15**: 1123–1141.

Lindley, D. V. 1964. The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35**: 1622–1643.

Lindsey, J. K., B. Jones, and A. F. Ebbutt. 1997. Simple models for repeated ordinal responses with an application to a seasonal rhinitis clinical trial. *Statist. Med.* **16**: 2873–2882.

Lindsey, P. J., and J. Kaufmann. 2004. Analysis of a longitudinal ordinal response clinical trial using dynamic models. *Appl. Statist.* **53**: 523–537.

Lipsitz, S. 1992. Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics* **48**: 271–281.

Lipsitz, S. R., and G. Fitzmaurice. 1996. The score test for independence in $R \times C$ contingency tables with missing data. *Biometrics* **52**: 751–762.

Lipsitz, S. R., K. Kim, and L. Zhao. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statist. Med.* **13**: 1149–1163.

Lipsitz, S. R., G. M. Fitzmaurice, and G. Molenberghs. 1996. Goodness-of-fit tests for ordinal response regression models. *Appl. Statist.* **45**: 175–190.

Little, R. J., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: Wiley.

Liu, I. 2003. Describing ordinal odds ratios for stratified $r \times c$ tables. *Biometrical J.* **45**: 730–750.

Liu, I.-M., and A. Agresti. 1996. Mantel–Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics* **52**: 1223–1234.

Liu, I., and A. Agresti. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments (with discussion). *Test* **14**: 1–73.

Liu, I., B. Mukherjee, T. Suesse, D. Sparrow, and S. K. Park. 2009. Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statist. Med.* **28**: 412–429.

Liu, L. C., and D. Hedeker. 2006. A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics* **62**, 261–268.

Lombardo, R., E. J. Beh, and L. D'Ambra. 2007. Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. *Comput. Statist. Data Anal.* **52**: 566–577.

Lovison, G. 2005. On Rao score and Pearson X^2 statistics in generalized linear models. *Statist. Papers* **46**: 555–574.

Lui, K.-J., X.-H. Zhou, and C.-D. Lin. 2004. Testing equality between two diagnostic procedures in paired-sample ordinal data. *Biometrical J.* **46**: 642–652.

Lumley, T. 1996. Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics* **52**: 354–361.

Lunn, D. J., J. Wakefield, and A. Racine-Poon. 2001. Cumulative logit models for ordinal data: A case study involving allergic rhinitis severity. *Statist. Med.* **20**: 2261–2285.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

Magidson, J. 1996. Maximum likelihood assessment of clinical trials based on an ordered categorical response. *Drug Inform. J.* **30**: 143–170.

Mantel, N. 1963. Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Assoc.* **58**: 690–700.

Mantel, N. 1979. Ridit analysis and related ranking procedures—Use at your own risk. *Amer. J. Epidemiol.* **109**: 25–29.

Mardia, K. V. 1967. Some contributions to contingency-type bivariate distributions. *Biometrika* **54**: 235–249.

Mark, S. D., and M. H. Gail. 1994. A comparison of likelihood-based and marginal estimating equation methods for analysing repeated ordered categorical responses with missing data. *Statist. Med.* **13**: 479–493.

Marshall, R. J. 1999. Classification to ordinal categories using a search partition methodology with an application in diabetes screening. *Statist. Med.* **18**: 2723–2735.

Martinson, E. O., and M. A. Hamdan. 1972. Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *J. Statist. Comput. Simul.* **1**: 45–54.

Masters, G. N. 1982. A Rasch model for partial credit scoring. *Psychometrika* **47**: 149–174.

Mayer, L. S. 1971. A note on treating ordinal data as interval data. *Amer. Sociol. Rev.* **36**: 519–520.

Mayer, L. S., and J. A. Robinson. 1978. Measures of association for multiple regression models with ordinal predictor variables. *Sociol. Methodol.* **9**: 141–163.

McCullagh, P. 1977. A logistic model for paired comparisons with ordered categorical data. *Biometrika* **64**: 449–453.

McCullagh, P. 1978. A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* **65**: 413–418.

McCullagh, P. 1980. Regression models for ordinal data. *J. Roy. Statist. Soc.* **B 42**: 109–142.

McCullagh, P. 1984. On the elimination of nuisance parameters in the proportional odds model. *J. Roy. Statist. Soc.* **B 46**: 250–256.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. (1st ed. 1983) London: Chapman & Hall.

McDonald, J. W., D. C. DeRoure, and D. T. Michaelides. 1998. Exact tests for two-way symmetric contingency tables. *Statist. Comput.* **8**: 391–399.

McKelvey, R. D., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **4**: 103–120.

Meeks, S. L., and R. B. D'Agostino. 1983. A model for comparisons with ordered categorical data. *Commun. Statist. Theory Methods* **12**: 895–906.

Mehta, C. R., N. R. Patel, and A. A. Tsiatis. 1984. Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40**: 819–825.

Mehta, C. R., N. Patel, and P. Senchaudhuri. 1992. Exact stratified linear rank test for ordered categorical and binary data. *J. Comput. Graph. Statist.* **1**: 21–40.

Miller, M. E., C. S. Davis, and J. R. Landis. 1993. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**: 1033–1044.

Miller, M. E., T. R. Ten Have, B. A. Reboussin, K. K. Lohman, and W. J. Rejeski. 2001. A marginal model for analyzing discrete outcomes from longitudinal survey with outcomes subject to multiple-cause nonresponse. *J. Amer. Statist. Assoc.* **96**: 844–857.

Min, Y., and A. Agresti 2005. Random effect models for repeated measures of zero-inflated count data. *Statist. Model.* **5**: 1–19.

Molenberghs, G., and E. Lesaffre. 1994. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.* **89**: 633–644.

Molenberghs, G., and E. Lesaffre. 1999. Marginal modeling of multivariate categorical data. *Statist. Med.* **18**: 2237–2255.

Molenberghs, G., and G. Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., and G. Verbeke. 2007. Likelihood ratio, score and Wald tests in a constrained parameter space. *Amer. Statist.* **61**: 22–27.

Molenberghs, G., M. G. Kenward, and E. Lesaffre. 1997. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**: 33–44.

Moustaki, I. 2000. A latent variable model for ordinal variables. *Appl. Psychol. Meas.* **24**: 211–233.

Moustaki, I. 2003. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *Brit. J. Math. Statist. Psychol.* **56**: 337–357.

Mukherjee, B., and I. Liu. 2008. A characterization of bias for fitting multivariate generalized linear models under choice-based sampling. *J. Multivariate Anal.* **100**: 459–472.

Mukherjee, B., I. Liu, and S. Sinha. 2007. Analysis of matched case–control data with multiple ordered disease states: Possible choices and comparisons. *Statist. Med.* **26**: 3240–3257.

Mukherjee, B., J. Ahn, I. Liu, P. Rathouz, and B. Sanchez. 2008. Fitting stratified proportional odds models by amalgamating conditional likelihoods. *Statist. Med.* **27**: 4950–4971.

Müller, G., and C. Czado. 2005. An autoregressive ordered probit model with application to high frequency financial data. *J. Comput. Graph. Statist.* **14**: 320–338.

Munzel, U., and L. A. Hothorn. 2001. A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. *Biometrical J.* **43**: 553–569.

Muthén, B. 1984. A general structural equation model with dichotomous, ordered categorical and continuous latent variables indicators. *Psychometrika* **49**: 115–132.

Mwalili, S. M., E. Lesaffre, and D. Declerck. 2004. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Appl. Statist.* **54**: 77–93.

Nair, V. N. 1986. Testing in industrial experiments with ordered categorical data. *Technometrics* **28**: 283–311.

Nair, V. N. 1987. Chi-squared-type tests for ordered alternatives in contingency tables. *J. Amer. Statist. Assoc.* **82**: 283–291.

Nandram, B. 1989. Discrimination between the complementary log-log and logistic model for ordinal data. *Commun. Statist. Theory Methods* **18**: 2155–2164.

Natarajan, R., and C. E. McCulloch. 1998. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *J. Comput. Graph. Statist.* **7**: 267–277.

Nguyen, T. T., and A. R. Sampson. 1987. Testing for positive quadrant dependence in ordinal contingency tables. *Naval Res. Logistics* **34**: 859–877.

Nores, M. L., and M. Diaz. 2008. Some properties of regression estimators in GEE models for clustered ordinal data. *Comput. Statist. Data Anal.* **52**: 3877–3888.

Ntzoufras, I. 2009. *Bayesian Modeling Using WinBUGS*. Hoboken, NJ: Wiley.

O'Connell, A. A. 2006. *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: Sage.

O'Connell, D. L., and A. J. Dobson. 1984. General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* **40**: 973–983.

O'Gorman, T. W., and R. F. Woolson. 1988. Analysis of ordered categorical data using the SAS system. In *Proc. 13th Annual SAS Users Group Conference*. Cary, NC: SAS Institute, pp. 957–963.

Oh, M. 1995. On maximum likelihood estimation of cell probabilities in $2 \times k$ contingency tables under negative dependence restrictions with various sampling schemes. *Commun. Statist. Theory Methods* **24**, 2127–2143.

O'Hagan, A., and J. Forster. 2004. *Kendall's Advanced Theory of Statistics: Bayesian Inference*. London: Edward Arnold.

Ohman-Strickland, P. A., and S.-E. Lu. 2003. Estimates, power and sample size calculations for two-sample ordinal outcomes under before–after study designs. *Statist. Med.* **22**: 1807–1818.

Olsson, U. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**: 443–460.

Olsson, U., F. Drasgow, and N. J. Dorans. 1982. The polyserial correlation coefficient. *Psychometrika* **47**: 337–347.

Oluyede, B. O. 1993. A modified chi-square test for testing equality of two multinomial populations against an ordering restricted alternative. *Biometrical J.* **35**: 997–1012.

Oluyede, B. O. 1994. A modified chi-square test of independence against a class of ordered alternatives in an $r \times c$ contingency table. *Canad. J. Statist.* **22**: 75–87.

Parsons, N. R., R. N. Edmondson, and S. G. Gilmour. 2006. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Appl. Statist.* **55**: 507–524.

Parsons, N. R., M. L. Costa, J. Achten, and N. Stallard. 2009. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Comput. Statist. Data Anal.* **53**: 632–641.

Patefield, W. M. 1982. Exact tests for trends in ordered contingency tables. *Appl. Statist.* **31**: 32–43.

Pearson, K. 1904. Mathematical contributions to the theory of evolution. XIII: On the theory of contingency and its relation to association and normal correlation. In *Draper's Co. Research Memoirs*. Biometric Series, no. 1. (Reprinted 1948 in *Karl Pearson's Early Papers*, ed. E. S. Pearson. Cambridge, UK: Cambridge University Press.)

Perkins, S. M., and M. P. Becker. 2002. Assessing rater agreement using marginal association models. *Statist. Med.* **21**: 1743–1760.

Perlman, M. D., and L. Wu. 1999. The emperor's new clothes. *Statist. Sci.* **14**: 355–369.

Peterson, B., and F. E. Harrell, Jr. 1990. Partial proportional odds models for ordinal response variables. *Appl. Statist.* **39**: 205–217.

Pettersson, T. 2002. A comparative study of model based tests of independence for ordinal data using the bootstrap. *J. Statist. Comput. Simul.* **72**: 187–203.

Pettitt, A. N. 1984a. Tied, grouped continuous and ordered categorical data: A comparison of two models. *Biometrika* **71**: 35–42.

Pettitt, A. N. 1984b. Proportional odds models for survival data and estimates using ranks. *Appl. Statist.* **33**: 169–175.

Piccarreta, R. 2008. Classification trees for ordinal variables. *Comput. Statist.* **23**: 407–427.

Plackett, R. L. 1965. A class of bivariate distributions. *J. Amer. Statist. Assoc.* **60**: 516–522.

Plackett, R. L., and S. R. Paul. 1978. Dirichlet models for square contingency tables. *Commun. Statist. Theory Methods* **7**: 939–952.

Plewis, I., F. Vitaro, and R. Tremblay. 2006. Modelling repeated ordinal reports from multiple informants. *Statist. Model.* **6**: 251–263.

Ploch, D. R. 1974. Ordinal measures of association and the general linear model. In *Measurement in the Social Sciences*, ed. H. M. Blalock. Hawthorne, NY: Aldine, Chap. 12.

Podgor, M. J., J. L. Gastwirth, and C. R. Mehta. 1996. Efficiency robust tests of independence in contingency tables with ordered classifications. *Statist. Med.* **15**: 2095–2105.

Pratt, J. W. 1981. Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76**: 103–106.

Prentice, R. L., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**: 57–67.

Pruscha, H. 1994. Partial residuals in cumulative regression models for ordinal data. *Statist. Papers* **35**: 273–284.

Pulkstenis, E., and T. J. Robinson. 2004. Goodness-of-fit test for ordinal response regression models. *Statist. Med.* **23**: 999–1014.

Qiu, Z., P. X.-K. Song, and M. Tan. 2002. Bayesian hierarchical models for multi-level repeated ordinal data using WinBUGS. *J. Biopharm. Statist.* **12**: 121–135.

Qu, Y. S., M. R. Piedmonte, and S. V. Medendor. 1995. Latent variable models for clustered ordinal data. *Biometrics* **51**: 268–275.

Qu, Y., and M. Tan. 1998. Analysis of clustered ordinal data with subclusters via a Bayesian hierarchical model. *Commun. Statist. Theory Methods* **27**: 1461–1476.

Quade, D. 1974. Nonparametric partial correlation. In *Measurement in the Social Sciences*, ed. H. M. Blalock. Hawthorne, NY: Aldine, Chap. 13.

Quinn, K. M. 2004. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Anal.* **12**: 338–353.

Rabbee, N., B. A. Coull, N. Patel, and P. Senchaudhuri. 2003. Power and sample size for ordered categorical data. *Statist. Methods Med. Res.* **12**: 73–84.

Rabe-Hesketh, S., and A. Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*, 2nd ed. Stata Press.

Rahlfs, V. W., and H. Zimmerman. 1993. Scores: Ordinal data with few categories—How should they be analyzed? *Drug Inform. J.* **27**: 1227–1240.

Raman, R., and D. Hedeker. 2005. A mixed-effects regression model for three-level ordinal response data. *Statist. Med.* **24**, 3331–3345.

Rampichini, C., L. Grilli, and A. Petrucci. 2004. Analysis of university course evaluations: From descriptive measures to multilevel models. *Statist. Methods Appl.* **13**: 357–373.

Rao, M. B., P. R. Krishnaiah, and K. Subramanyam. 1987. A structure theorem on bivariate positive quadrant dependent distributions and test for independence in two-way contingency tables. *J. Multivariate Anal.* **23**: 93–118.

Rayner, J. C. W., and D. J. Best. 2000. A smooth analysis of singly ordered two-way contingency tables. *J. Appl. Math. Decision Sci.* **4**: 83–98.

Rayner, J. C. W., and D. J. Best. 2001. *A Contingency Table Approach to Nonparametric Testing*. London: Chapman & Hall.

Ribaudo, H. J., J. Bernhard, M. Bacchi, and S. G. Thompson. 1999. A multilevel analysis of longitudinal ordinal data: Evaluation of the level of physical performance of women receiving adjuvant therapy for breast cancer. *J. Roy. Statist. Soc.* A **162**: 349–360.

Ritchie-Scott, A. 1918. The correlation coefficient of a polychoric table. *Biometrika* **12**: 93–133.

Ritov, Y., and Z. Gilula. 1991. The order-restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Statist.* **19**: 2090–2101.

Ritov, Y., and Z. Gilula. 1993. Analysis of contingency tables by correspondence models subject to order constraints. *J. Amer. Statist. Assoc.* **88**: 1380–1387.

Roberts, C., and R. McNamee. 2005. Assessing the reliability of ordered categorical scales using kappa-type statistics. *Statist. Methods Med. Res.* **14**: 493–514.

Robertson, T. 1978. Testing for and against an order restriction on multinomial parameters. *J. Amer. Statist. Assoc.* **73**: 197–202.

Robertson, T. and F. T. Wright. 1981. Likelihood-ratio tests for and against a stochastic ordering between multinomial populations. *Ann. Statist.* **9**: 1248–1257.

Robertson, T., F. T. Wright, and R. L. Dykstra. 1988. *Order Restricted Statistical Inference*. New York: Wiley.

Rom, D., and S. K. Sarkar. 1992. A generalized model for the analysis of association in ordinal contingency tables. *J. Statist. Plann. Inference* **33**: 205–212.

Ronning, G., and M. Kukuk. 1996. Efficient estimation of ordered probit model. *J. Amer. Statist. Assoc.* **91**: 1120–1129.

Rosenthal, I. 1966. Distribution of the sample version of the measure of association, gamma. *J. Amer. Statist. Assoc.* **61**: 440–453.

Rossi, P. E., Z. Gilula, and G. M. Allenby. 2001. Overcoming scale usage heterogeneity: A Bayes hierarchical approach. *J. Amer. Statist. Assoc.* **96**: 20–31.

Rossini, A. J., and A. A. Tsiatis. 1996. A semiparametric proportional odds regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91**: 713–721.

Rudolfer, S. M., P. C. Watson, and E. Lesaffre. 1995. Are ordinal models useful for classification? A revised analysis. *J. Statist. Comput. Simul.* **52**: 105–132.

Ryan, L. 1992. Quantitative risk assessment for developmental toxicity. *Biometrics* **48**: 163–174.

Ryu, E. 2009. Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. *Statist. Med.* **29**: 3179–3188.

Ryu, E., and A. Agresti 2008. Modeling and inference for an ordinal effect size measure. *Statist. Med.* **27**: 1703–1717.

Saei, A., J. Ward, and C. A. McGilchrist. 1996. Threshold models in a methodone programme evaluation. *Statist. Med.* **15**: 2253–2260.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* **17**.

Sarkar, S. K. 1989. Quasi-independence in ordinal triangular contingency tables. *J. Amer. Statist. Assoc.* **84**: 592–597.

Schriever, B. F. 1983. Scaling of order dependent categorical variables with correspondence analysis. *Internat. Statist. Rev.* **51**: 225–238.

Schuster, C., and A. von Eye. 2001. Models for ordinal agreement data. *Biometrical J.* **43**: 795–808.

Scott, S. C., M. S. Goldberg, and N. E. Mayo. 1997. Statistical assessment of ordinal outcomes in comparative studies. *J. Clin. Edidemiol.* **50**: 45–55.

Sedransk, J., J. Monahan, and H. Y. Chiu. 1985. Bayesian estimation of finite population parameters in categorical data models incorporating order. *J. Roy. Statist. Soc.* **B 47**: 519–527.

Semenya, K., G. G. Koch, M. E. Stokes, and R. N. Forthofer. 1983. Linear models methods for some rank function analyses of ordinal categorical data. *Commun. Statist. Theory Methods* **12**: 1277–1298.

Senn, S. 2007. Drawbacks to noninteger scoring for ordered categorical data. *Biometrics* **63**: 296–298.

Shah, D. A., and L. V. Madden. 2004. Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology* **94**: 33–43.

Shashua, A., and A. Levin. 2003. Ranking with large margin principle: Two approaches. *Adv. Neural Inform. Process. Syst.* **15**. 937–944.

Siciliano, R., and A. Mooijaart. 1997. Three-factor association models for three-way contingency tables. *Comput. Statist. Data Anal.* **24**: 337–356.

Silvapulle, M. J., and P. K. Sen. 2004. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Hoboken, NJ: Wiley.

Simon, G. 1974. Alternative analyses for the singly-ordered contingency table. *J. Amer. Statist. Assoc.* **69**: 971–976.

Simon, G. 1978. Efficacies of measures of association for ordinal contingency tables. *J. Amer. Statist. Assoc.* **73**: 545–551.

Simonoff, J. 1987. Probability estimation via smoothing in sparse contingency tables with ordered categories. *Statist. Probab. Lett.* **5**: 55–63.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Simonoff, J. S. 1998. Three sides of smoothing: Categorical data smoothing, nonparametric regression, and density estimation. *Internat. Statist. Rev.* **66**: 137–156.

Simonoff, J. S., and C.-L. Tsai. 1991. Higher-order effects in log-linear and log-non-linear models for contingency tables with ordered categories. *Appl. Statist.* **40**: 449–458.

Simonoff, J. S., Y. Hochberg, and B. Reiser. 1986. Alternative estimation procedures for Pr($X < Y$) in categorized data. *Biometrics* **42**: 895–907.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* London: Chapman & Hall.

Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica* **55**: 409–424.

Smith, P. J., T. J. Thompson, M. M. Engelgau, and W. H. Herman. 1996. A generalized linear model for analysing receiver operating characteristic curves. *Statist. Med.* **15**: 323–333.

Smith, R. B. 1974. Continuities in ordinal path analysis. *Social Forces* **53**: 200–229.

Smyth, G. K. 2003. Pearson's goodness of fit statistic as a score test statistic. In *Science and Statistics: A Festschrift for Terry Speed*, ed. D. R. Goldstein. IMS Lecture Notes–Monograph Series, vol. **40**. Hayward, CA: Institute of Mathematical Statistics, pp. 115–126.

Snapinn, S. M., and R. D. Small. 1986. Tests of significance using regression models for ordered categorical data. *Biometrics* **42**: 583–592.

Snell, E. J. 1964. A scaling procedure for ordered categorical data. *Biometrics* **20**: 592–607.

Sobel, M. E. 1988. Some models for the multiway contingency table with a one-to-one correspondence among categories. *Sociol. Methodol.* **18**: 165–192.

Sobel, M. E. 1997. Modeling symmetry, asymmetry, and change in ordered scales with midpoints using adjacent category logit models for discrete data. *Sociol. Methods Res.* **26**: 213–232.

Sobel, M. E., M. P. Becker, and S. M. Minick. 1998. Origins, destinations, and association in occupational mobility. *Amer. J. Sociol.* **104**: 687–721.

Somers, R. H. 1959. The rank analogue of product moment partial correlation and regression, with application to manifold, ordered contingency tables. *Biometrika* **46**: 241–246.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *Amer. Sociol. Rev.* **27**: 799–811.

Somers, R. H. 1968. An approach to the multivariate analysis of ordinal data. *Amer. Sociol. Rev.* **33**: 971–977.

Spitzer, R. L., J. Cohen, J. L. Fleiss, and J. Endicott. 1967. Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry* **17**: 83–87.

StatXact. 2005. *StatXact 7: Statistical Software for Exact Nonparametric Inference*, vol. 1 and 2. Cytel Software.

Steele, F., and H. Goldstein. 2006. A multilevel factor model for mixed binary and ordinal indicators of women's status. *Sociol. Methods Res.* **35**: 137–153.

Stiger, T. R., H. X. Barnhart, and J. M. Williamson. 1999. Testing proportionality in the proportional odds model fitted with GEE. *Statist. Med.* **18**: 1419–1433.

Stokes, M. E., C. S. Davis, and G. G. Koch. 2000. *Categorical Data Analysis Using the SAS System*, 2nd ed. Cary, NC: SAS Institute.

Stram, D. O., L. J. Wei, and J. H. Ware. 1988. Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *J. Amer. Statist. Assoc.* **83**: 631–637.

Streitberg, B., and J. Röhmel. 1988. Simultane Anwendung aller möglichen Tests zu einem vorgelegten Test Problem. In *Multiple Hypothesenprüfung*, ed. P. Bauer, G. Hommel, and E. Sonnemann. Berlin: Springer-Verlag.

Stuart, A. 1953. The estimation and comparison of strengths of association in contingency tables. *Biometrika* **40**: 105–110.

Stuart, A. 1963. Calculation of Spearman's rho for ordered two-way classifications. *Amer. Statist.* **17**: 23–24.

Svensson, E. 1997. A coefficient of agreement adjusted for bias in paired ordered categorical data. *Biometrical J.* **39**: 643–657.

Svensson, E. 1998. Ordinal invariant measures for individual and group changes in ordered categorical data. *Statist. Med.* **17**: 2923–2936.

Svensson, E. 2000a. Concordance between ratings using different scales for the same variable. *Statist. Med.* **19**: 3483–3496.

Svensson, E. 2000b. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical J.* **42**: 417–434.

Svensson, E., and S. Holm. 1994. Separation of systematic and random differences in ordinal rating scales. *Statist. Med.* **13**: 2437–2453.

Taguchi, G. 1974. A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. *Saishin Igaku [The Newest Medicine]* **29**: 806–813.

Takeuchi, K. and C. Hirotsu. 1982. The cumulative chi-squares method against ordered alternatives in two-way contingency tables. *Rep. Statist. Appl. Res.* **29**: 1–13.

Tallis, G. 1962. The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18**: 342–353.

Tan, M., Y. S. Qu, E. Mascha, and A. Schubert. 1999. A Bayesian hierarchical model for multi-level repeated ordinal data: analysis of oral practice examinations in a large anaesthesiology training program. *Statist. Med.* **18**: 1983–1992.

Tarantola, C., G. Consonni, and P. Dellaportas. 2008. Bayesian clustering for row effects models. *J. Statist. Plann. Inference* **138**: 2223–2235.

Ten Have, T. R. 1996. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **52**: 473–491.

Ten Have, T. R., and D. H. Uttal. 1994. Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Appl. Statist.* **43**: 371–384.

Ten Have, T., J. R. Landis, and J. Hartzel. 1998. Population-averaged and cluster-specific models for clustered ordinal response data. *Statist. Med.* **15**: 2573–2588.

Ten Have, T. R., M. E. Miller, B. A. Reboussin, and M. M. James. 2000. Mixed effects logistic regression models for longitudinal ordinal functional response data with multiple cause drop-out from the longitudinal study of aging. *Biometrics* **56**: 279–287.

Terza, J. V. 1985. Ordinal probit: A generalization. *Commun Statist. Theory Methods* **14**: 1–11.

Thompson, W. A. 1977. On the treatment of grouped observations in life studies. *Biometrics* **33**: 463–470.

Thompson, R., and R. J. Baker. 1981. Composite link functions in generalized linear models. *Appl. Statist.* **30**: 125–131.

Titterington, D. M., and A. W. Bowman. 1985. A comparative study of smoothing procedures for ordered categorical data. *J. Statist. Comput. Simul.* **21**: 291–312.

Todem, D., K. Kim, and E. Lesaffre. 2007. Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statist. Med.* **26**: 1034–1054.

Toledano, A., and C. Gatsonis. 1996. Ordinal regression methodology for ROC curves derived from correlated data. *Statist. Med.* **15**: 1807–1826.

Toledano, A., and C. Gatsonis. 1999. Generalized estimating equations for ordinal categorical data: Arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* **55**: 488–496.

Tosteson, A. N. A., and C. B. Begg. 1988. A general regression methodology for ROC curve estimation. *Med. Decision Making* **8**: 204–215.

Tosteson, T. D., L. A. Stefanski, and D. W. Schafer. 1989. A measurement-error model for binary and ordinal regression. *Statist. Med.* **8**: 1139–1147.

Tosteson, A. N., M. C. Weinstein, J. Wittenberg, and C. B. Begg. 1994. ROC curve regression analysis: The use of ordinal regression models for diagnostic test assessment. *Environ. Health Perspect.* **102**: 73–78.

Tsai, M.-T., and P. K. Sen. 1995. A test of quasi-independence in ordinal triangular contingency tables. *Statist. Sinica* **5**: 767–780.

Tutz, G. 1986. Bradley–Terry–Luce models with an ordered response. *J. Math. Psychol.* **30**: 306–316.

Tutz, G. 1989. Compound regression models for ordered categorical data. *Biometrical J.* **31**: 259–272.

Tutz, G. 1990. Sequential item response models with an ordered response. *Brit. J. Math. Statist. Psychol.* **43**: 39–55.

Tutz, G. 1991. Sequential models in categorical regression. *Comput. Statist. Data Anal.* **11**: 275–295.

Tutz, G. 2003. Generalized semiparametrically structured ordinal models. *Biometrics* **59**: 263–273.

Tutz, G., and H. Binder. 2004. Flexible modelling of discrete failure time including time-varying smooth effects. *Statist. Med.* **23**: 2445–2461.

Tutz, G., and K. Hechenbichler. 2005. Aggregating classifiers with ordinal response structure. *J. Statist. Comput. Simul.* **75**: 391–408.

Tutz, G., and W. Hennevogl. 1996. Random effects in ordinal regression models. *Comput. Statist. Data Anal.* **22**: 537–557.

Uebersax, J. S. 1993. Statistical modeling of expert ratings on medical treatment appropriateness. *J. Amer. Statist. Assoc.* **88**: 421–427.

Uebersax, J. S. 1999. Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Appl. Psychol. Meas.* **23**: 283–297.

Uebersax, J. S., and W. M. Grove. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* **49**: 823–835.

Valet, F., C. Guinot, and J. Y. Mary. 2007. Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. *Statist. Med.* **26**: 647–662.

van der Heijden, P. G. M., A. de Falguerolles, and J. de Leeuw. 1989. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Statist.* **38**: 249–292.

Vargha, A., and H. D. Delaney. 1998. The Kruskal–Wallis test and stochastic homogeneity. *J. Educ. Behav. Statist.* **59**: 137–142.

Varin, C., and C. Czado. 2010. A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11**: 127–138.

Varin, C., and P. Vidoni. 2006. Pairwise likelihood inference for ordinal categorical time series. *Comput. Statist. Data Anal.* **51**: 2365–2373.

Vermunt, J. K. 1999. A general class of nonparametric models for ordinal categorical data. *Sociol. Methodol.* **29**: 187–223.

Vermunt, J. K. 2001. The use of restricted latent class models for defining and testing nonparametric item response theory models. *Appl. Psychol. Meas.* **25**: 283–294.

Vermunt, J. K., M. F. Rodrigo, and M. Ato-Garcia. 2001. Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociol. Methods Res.* **30**: 170–196.

Vigderhous, G. 1979. Equivalence between ordinal measures of association and tests of significant differences between samples. *Quality Quantity* **13**: 187–201.

Vijn, P. 1983. Ordinal data, ordered scale points, and order statistics. *Psychometrika* **48**: 437–449.

von Eye, A., and E. Y. Mun. 2005. *Analyzing Rater Agreement: Manifest Variable Methods*. Mahwah, NJ: Lawrence Erlbaum.

Waegemana, W., B. De Baets, and L. Boullart. 2008. ROC analysis in ordinal regression learning. *Pattern Recognit. Lett.* **29**: 1–9.

Wahrendorf, J. 1980. Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. *Biometrika* **67**: 15–21.

Walker, S. H., and D. B. Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**: 167–179.

Wang, Y. 1996. A likelihood ratio test against stochastic ordering in several populations. *J. Amer. Statist. Assoc.* **91**: 1676–1683.

Wang, Y. J. 1987. The probability integrals of bivariate normal distributions: A contingency table approach. *Biometrika* **74**: 185–190.

Wang, Y. J. 1997. Multivariate normal integrals and contingency tables with ordered categories. *Psychometrika* **62**: 267–284.

Webb, E. L., and J. J. Forster. 2008. Bayesian model determination for multivariate ordinal and binary data. *Comput. Statist. Data Anal.* **52**: 2632–2649.

Wei, L. J., and J. M. Lachin. 1984. Distribution-free tests for incomplete multivariate observations. *J. Amer. Statist. Assoc.* **79**: 653–661.

Weisberg, H. I. 1972. Bayesian comparison of two ordered multinomial populations. *Biometrics* **28**: 859–867.

Weiss, A. A. 1993. A bivariate ordered probit model with truncation. *Appl. Statist.* **42**: 487–499.

Wermuth, N., and D. R. Cox. 1998. On the application of conditional independence to ordinal data. *Internat. Statist. Rev.* **66**: 181–199.

White, A. A., J. R. Landis, and M. M. Cooper. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Internat. Statist. Rev.* **50**: 27–34.

Whitehead, J. 1993. Sample size calculations for ordered categorical data. *Statist. Med.* **12**: 2257–2271.

Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika* **39**: 274–289.

Williams, O. D., and J. E. Grizzle. 1972. Analysis for contingency tables having ordered response categories. *J. Amer. Statist. Assoc.* **67**: 55–63.

Williams, R. 2009. Using heterogeneous choice models to compare logit and probit coefficient across groups. *Sociol. Methods Res.* **37**: 531–559.

Williamson, J., and K. Kim. 1996. A global odds ratio regression model for bivariate ordered categorical data from opthalmologic studies. *Statist. Med.* **15**: 1507–1518.

Williamson, J., and M.-L. T. Lee. 1996. A GEE regression model for the association between an ordinal and a nominal variable. *Commun. Statist. Theory Methods* **25**: 1887–1901.

Williamson, J., and A. K. Manatunga. 1997. Assessing interrater agreement from dependent data. *Biometrics* **53**: 707–714.

Williamson, J., K. Kim, and S. Lipsitz. 1995. Analyzing bivariate ordinal data using a global odds ratio. *J. Amer. Statist. Assoc.* **90**: 1432–1437.

Williamson, J., S. R. Lipsitz, and K. Kim. 1999. GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data. *Comput. Methods Prog. Biomed.* **58**: 25–34.

Winship, C., and R. D. Mare. 1984. Regression models with ordinal variables. *Amer. Sociol. Rev.* **49**: 512–525.

Wolfe, R., and D. Firth. 2002. Modelling subjective use of an ordinal response scale in a many period crossover experiment. *J. Roy. Statist. Soc.* C **51**: 245–255.

Wong, R. S.-K. 2001. Multidimensional association models: A multilinear approach. *Sociol. Methods Res.* **30**: 197–240.

Xiang, L., K. K. W. Yau, and S. K. Tse. 2008. Influence diagnostics for stratified ordinal contingency tables. *J. Statist. Comput. Simul.* **78**: 405–415.

Xie, M., D. G. Simpson, and R. J. Carroll. 2000. Random effects in censored ordinal regression: Latent structure and Bayesian approach. *Biometrics* **56**: 376–383.

Xie, Y. 1992. The log-multiplicative layer models for comparing mobility tables. *Amer. Sociol. Rev.* **57**: 380–395.

Yamaguchi, K. 1990. Homophily and social distance in the choice of multiple friends. *J. Amer. Statist. Assoc.* **85**: 356–366.

Yates, F. 1948. The analysis of contingency tables with grouping based on quantitative characters. *Biometrika* **35**: 176–181.

Yee, T. W. 2010. The VGAM package for categorical data analysis. *J. Statist. Software*, to appear.

Yee, T. W., and T. J. Hastie. 2003. Reduced-rank vector generalized linear models. *Statist. Model.* **3**: 15–41.

Yee, T. W., and C. J. Wild. 1996. Vector generalized additive models. *J. Roy. Statist. Soc.* B **58**: 481–493.

Zaslavsky, A., and E. T. Bradlow. 1999. A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Amer. Statist. Assoc.* **94**: 43–52.

Zayeri, F., A. Kazemnejad, N. Khanafshar, and F. Nayeri. 2005. Modeling repeated ordinal responses using a family of power transformations: Application to neonatal hypothermia. *BMC Med. Res. Methodol.* **5**: 29.

Zheng, G. 2008. Analysis of ordered categorical data: Two score-independent approaches. *Biometrics* **64**: 1276–1279.

Example Index

Subject Index

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Jozef L. Teugels*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

CLARKE and DISNEY · Probability and Random Processes: A First Course with
   Applications, *Second Edition*
* COCHRAN and COX · Experimental Designs, *Second Edition*
   COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications
   in the Social, Behavioral, and Health Sciences
   CONGDON · Applied Bayesian Modelling
   CONGDON · Bayesian Models for Categorical Data
   CONGDON · Bayesian Statistical Modelling
   CONOVER · Practical Nonparametric Statistics, *Third Edition*
   COOK · Regression Graphics
   COOK and WEISBERG · Applied Regression Including Computing and Graphics
   COOK and WEISBERG · An Introduction to Regression Graphics
   CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture
   Data, *Third Edition*
   COVER and THOMAS · Elements of Information Theory
   COX · A Handbook of Introductory Statistical Methods
* COX · Planning of Experiments
   CRESSIE · Statistics for Spatial Data, *Revised Edition*
   CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
   DANIEL · Applications of Statistics to Industrial Experimentation
   DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data,
      Second Edition
   DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
   DAVID and NAGARAJA · Order Statistics, *Third Edition*
* DEGROOT, FIENBERG, and KADANE · Statistics and the Law
   DEL CASTILLO · Statistical Process Adjustment for Quality Control
   DeMARIS · Regression with Social Data: Modeling Continuous and Limited Response
   Variables
   DEMIDENKO · Mixed Models: Theory and Applications
   DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear
   Classification and Regression
   DETTE and STUDDEN · The Theory of Canonical Moments with Applications in
   Statistics, Probability, and Analysis
   DEY and MUKERJEE · Fractional Factorial Plans
   DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
   DODGE · Alternative Methods of Regression
* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
* DOOB · Stochastic Processes
   DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
   DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
   DRYDEN and MARDIA · Statistical Shape Analysis
   DUDEWICZ and MISHRA · Modern Mathematical Statistics
   DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences,
      Third Edition
   DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
   EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human
   Health Risk Assessment
* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
   ENDERS · Applied Econometric Time Series
† ETHIER and KURTZ · Markov Processes: Characterization and Convergence
   EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

HOEL · Introduction to Mathematical Statistics, *Fifth Edition*

HOGG and KLUGMAN · Loss Distributions

HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*

HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*

HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*

† HUBER and RONCHETTI · Robust Statistics, *Second Edition*

HUBERTY · Applied Discriminant Analysis

HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*

HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek— with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

IMAN and CONOVER · A Modern Approach to Statistics

† JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*

JOHNSON and KOTZ · Distributions in Statistics

JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions

JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*

JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*

JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations

JUREK and MASON · Operator-Limit Distributions in Probability Theory

KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence

KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*

KARIYA and KURATA · Generalized Least Squares

KASS and VOS · Geometrical Foundations of Asymptotic Inference

† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

KEDEM and FOKIANOS · Regression Models for Time Series Analysis

KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*

KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models

KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*

KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*

KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications

KOWALSKI and TU · Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

KROONENBERG · Applied Multiway Data Analysis

KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering

LACHIN · Biostatistical Methods: The Assessment of Relative Risks

LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*

LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry

LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*

LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*

LAWSON · Statistical Methods in Spatial Epidemiology

LE · Applied Categorical Data Analysis

LE · Applied Survival Analysis

LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*

LePAGE and BILLARD · Exploring the Limits of Bootstrap

LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics

LIAO · Statistical Group Comparison

LINDVALL · Lectures on the Coupling Method

LIN · Introductory Stochastic Analysis for Finance and Insurance

LINHART and ZUCCHINI · Model Selection

LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*

LLOYD · The Statistical Analysis of Categorical Data

LOWEN and TEICH · Fractal-Based Point Processes

MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*

MALLER and ZHOU · Survival Analysis with Long Term Survivors

MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel

MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data

MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets

MARCHETTE · Random Graphs for Statistical Pattern Recognition

MARDIA and JUPP · Directional Statistics

MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*

McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*

McFADDEN · Management of Data in Clinical Trials, *Second Edition*

* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition

McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data

McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*

McLACHLAN and PEEL · Finite Mixture Models

McNEIL · Epidemiological Research Methods

MEEKER and ESCOBAR · Statistical Methods for Reliability Data

MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice

MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

Regression, *Third Edition*
* MILLER · Survival Analysis, *Second Edition*
  MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
  MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fourth Edition*
  MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
  MUIRHEAD · Aspects of Multivariate Statistical Theory
  MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
  MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
  MURTHY, XIE, and JIANG · Weibull Models
  MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
  MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
† NELSON · Applied Life Data Analysis
  NEWMAN · Biostatistical Methods in Epidemiology
  OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
  OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
  OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
  PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
  PANJER · Operational Risk: Modeling and Analytics
  PANKRATZ · Forecasting with Dynamic Regression Models
  PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
* PARZEN · Modern Probability Theory and Its Applications
  PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
  PIANTADOSI · Clinical Trials: A Methodologic Perspective
  PORT · Theoretical Probability for Applications
  POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
  POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality
  PRESS · Bayesian Statistics: Principles, Models, and Applications
  PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
  PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
  PUKELSHEIM · Optimal Experimental Design
  PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
  QIU · Image Processing and Jump Regression Analysis
* RAO · Linear Statistical Inference and Its Applications, *Second Edition*
  RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
  RENCHER · Linear Models in Statistics
  RENCHER · Methods of Multivariate Analysis, *Second Edition*
  RENCHER · Multivariate Statistical Inference with Applications
* RIPLEY · Spatial Statistics
* RIPLEY · Stochastic Simulation
  ROBINSON · Practical Strategies for Experimenting
  ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*

ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance

ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice

ROSS · Introduction to Probability and Statistics for Engineers and Scientists

ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing

† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection

* RUBIN · Multiple Imputation for Nonresponse in Surveys

RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*

RUBINSTEIN and MELAMED · Modern Simulation and Modeling

RYAN · Modern Engineering Statistics

RYAN · Modern Experimental Design

RYAN · Modern Regression Methods, *Second Edition*

RYAN · Statistical Methods for Quality Improvement, *Second Edition*

SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications

* SCHEFFE · The Analysis of Variance

SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application

SCHOTT · Matrix Analysis for Statistics, *Second Edition*

SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives

SCHUSS · Theory and Applications of Stochastic Differential Equations

SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization

† SEARLE · Linear Models for Unbalanced Data

† SEARLE · Matrix Algebra Useful for Statistics

† SEARLE, CASELLA, and McCULLOCH · Variance Components

SEARLE and WILLETT · Matrix Algebra for Applied Economics

SEBER · A Matrix Handbook For Statisticians

† SEBER · Multivariate Observations

SEBER and LEE · Linear Regression Analysis, *Second Edition*

† SEBER and WILD · Nonlinear Regression

SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems

* SERFLING · Approximation Theorems of Mathematical Statistics

SHAFER and VOVK · Probability and Finance: It's Only a Game!

SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions

SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference

SRIVASTAVA · Methods of Multivariate Statistics

STAPLETON · Linear Statistical Models, *Second Edition*

STAPLETON · Models for Probability and Statistical Inference: Theory and Applications

STAUDTE and SHEATHER · Robust Estimation and Testing

STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*

STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics

STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods

STYAN · The Collected Papers of T. W. Anderson: 1943–1985

SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research

TAKEZAWA · Introduction to Nonparametric Regression

TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications

TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory

THOMPSON · Empirical Model Building

THOMPSON · Sampling, *Second Edition*

THOMPSON · Simulation: A Modeler's Approach

THOMPSON and SEBER · Adaptive Sampling

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets

TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness

TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics

TSAY · Analysis of Financial Time Series, *Second Edition*

UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data

† VAN BELLE · Statistical Rules of Thumb, *Second Edition*

VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*

VESTRUP · The Theory of Measures and Integration

VIDAKOVIC · Statistical Modeling by Wavelets

VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments

WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data

WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models

WEISBERG · Applied Linear Regression, *Third Edition*

WELSH · Aspects of Statistical Inference

WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment

WHITTAKER · Graphical Models in Applied Multivariate Statistics

WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting

WONNACOTT and WONNACOTT · Econometrics, *Second Edition*

WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles

WOODWORTH · Biostatistics: A Bayesian Introduction

WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*

WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*

WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis

YANG · The Construction Theory of Denumerable Markov Processes

YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics

ZACKS · Stage-Wise Adaptive Designs

ZELTERMAN · Discrete Distributions—Applications in the Health Sciences

* ZELLNER · An Introduction to Bayesian Inference in Econometrics

ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine