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

Analysis of Ordinal Categorical Data, Second Edition, By Alan Agresti

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:

$$y = 1 \text{ if } y^* \leq 20, \quad y = 2 \text{ if } 20 < y^* \leq 40, \quad y = 3 \text{ if } 40 < y^* \leq 60,$$
$$y = 4 \text{ if } 60 < y^* \leq 80, \quad y = 5 \text{ if } y^* > 80.$$

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

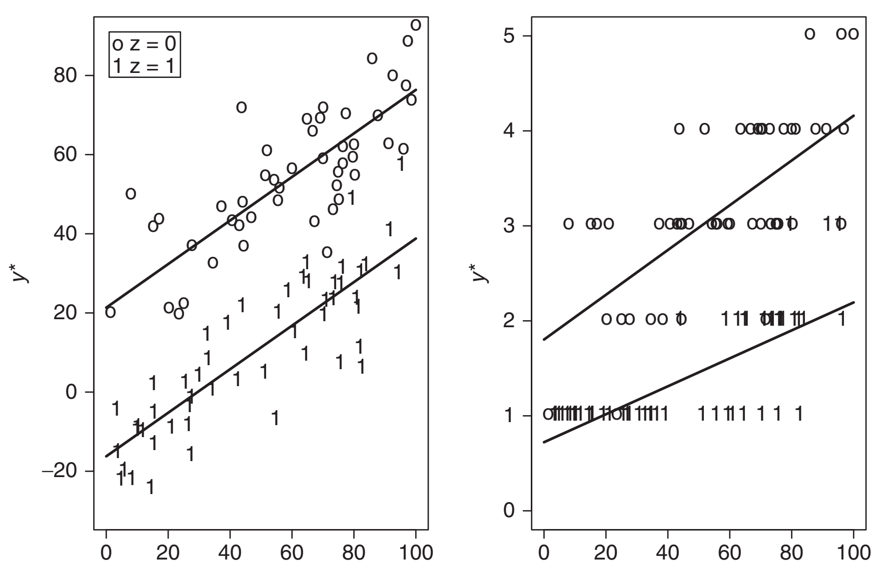

Figure 1.1. Ordered categorical data (in second panel) for which ordinary regression suggests interaction, because of a floor effect, but ordinal modeling does not. The data were generated (in first panel) from a normal main-effects regression model with continuous (x) and binary (z) explanatory variables. 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).