Modeling Ordered Choices

A PRIMER

William H. Greene
and David A. Hensher

CAMBRIDGE

Preface

This book began as a short note to propose the estimator in Section 8.3. In researching the recent developments in ordered choice modeling, we concluded that it would be useful to include some pedagogical material about uses and interpretation of the model at the most basic level. Our review of the literature revealed an impressive breadth and depth of applications of ordered choice modeling, but no single source that provided a comprehensive summary. There are several somewhat narrow surveys of the basic ordered probit/logit model, including Winship and Mare (1984), Becker and Kennedy (1992), Daykin and Moffatt (2002) and Boes and Winkelmann (2006a), and a book-length treatment, by Johnson and Albert (1999) that is focused on Bayesian estimation of the basic model parameters using grouped data. (See, also, Congdon (2005), Ch. 7 and Agresti (2002), Section 7.4.) However, these stop well short of examining the extensive range of variants of the model and the variety of fields of applications, such as bivariate and multivariate models, two-part models, duration models, panel data models, models with anchoring vignettes, semiparametric approaches, and so on. (We have, of necessity, omitted mention of many – perhaps most – of the huge number of applications.) This motivated us to assemble this more complete overview of the topic. As this review proceeded, it struck us that a more thorough survey of the model itself, including its historical development, might also be useful and (we hope) interesting for readers. The following is also a survey of the methodological literature on modeling ordered outcomes and ordered choices.

The development of the ordered choice regression model has emerged in two surprisingly disjointed strands of literature: in its earliest forms in the bioassay literature, and in its modern social science counterpart with the pioneering paper by McKelvey and Zavoina (1975) and its successors, such as Terza (1985). There are a few prominent links between these two literatures, notably Walker and Duncan (1967). However, even up to the contemporary literature, biological scientists and social scientists have largely successfully avoided bumping into each other. For example, the 500+ entry references list of this survey shares only four items with its 100+ entry counterpart in Johnson and Albert (1999).

The earliest applications of modeling ordered outcomes involved aggregate (grouped) data assembled in table format, and with moderate numbers of levels of usually a single stimulus. The fundamental ordered logistic ("cumulative odds") model in its various forms serves well as an appropriate modeling framework for such data. Walker and Duncan (1967) focused on a major limitation of the approach. When data are obtained with large numbers of inputs – the models in Brewer *et al.* (2008), for example, involve over forty covariates – and many levels of those inputs, then cross-tabulations are no longer feasible or adequate. Two requirements become obvious: the use of the individual data and the heavy reliance on what amount to multiple regression-style techniques. McKelvey and Zavoina (1975) added to the model a reliance on a formal underlying "data-generating process," the latent regression. This mechanism makes an occasional appearance in the bioassay treatment, but is never absent from the social science application. The *cumulative odds model* for contingency tables and the fundamental *ordered probit model* for individual data are now standard tools. The recent advances in ordered choice modeling have involved modeling heterogeneity, in cross-sections and in panel data sets. These include a variety of threshold models and models of parameter variation such as latent class and mixed and hierarchical models. The chapters in this book present, in some detail, the full range of varieties of models for ordered choices.

This book is intended to be a survey of a particular class of discrete choice models. We anticipate that it can be used in a graduate level course in applied econometrics or statistics at the level of, say, Greene (2008a) or Wooldridge (2002b) and as a reference in specialized courses such as microeconometrics or discrete choice modeling. We assume that the reader is familiar with basic statistics and econometrics and with modeling techniques somewhat beyond the linear regression model. An introduction to maximum likelihood estimation and the most familiar binary choice models, probit and logit, is assumed, though developed in great detail in Chapter 2. The focus of this book is on areas of application of ordered choice models. The range of applications considered here includes economics, sociology, health economics, finance, political science, statistics in medicine, transportation planning, and many others. We have drawn on all of these in our collection of applications. We leave it to others, e.g., Hayashi (2000), Wooldridge (2002a), or Greene (2008a) to provide background material on, e.g., asymptotic theory for estimators and practical aspects of nonlinear optimization.

All of the computations carried out here were done with *NLOGIT* (see www.nlogit.com). Most of them can also be done with several other packages, such as *Stata* and SAS. Since this book is not a "how to" guide for any particular computer program, we have not provided any instructions on how to obtain the results with *NLOGIT* (or any other program). We assume that the interested reader can follow through on our developments with their favorite software, whatever that might be. Rather, our interest is in the models and techniques.

We would like to thank Joseph Hilbe and Chandra Bhat for their suggestions that have improved this work and Allison Greene for her assistance with the manuscript. Any errors that remain are ours.

Contents

Tables

Figures

1 Introduction: random utility and ordered choice models

Netflix (www.netflix.com) is an internet company that rents movies on DVDs to subscribers. The business model works by having subscribers order the DVD online for home delivery and return by regular mail. After a customer returns a DVD, the next time they log on to the website, they are invited to rate the movie on a five-point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who rented that movie are averaged to provide a recommendation to prospective viewers. For example, as of April 5, 2009, the average rating of the 2007 movie *National Treasure: Book of Secrets* given by approximately 12,900 visitors to the site was 3.8. This rating process provides a natural application of the models and methods that interest us in this book.

For any individual viewer, we might reasonably hypothesize that there is a continuously varying strength of preferences for the movie that would underlie the rating they submit. For convenience and consistency with what follows, we will label that strength of preference "utility," U^*. Given that there are no natural units of measurement, we can describe utility as ranging over the entire real line:

$$-\infty < U^*_{im} < +\infty$$

where i indicates the individual and m indicates the movie. Individuals are invited to "rate" the movie on an integer scale from one to five. Logically, then, the translation from underlying utility to a rating could be viewed as a *censoring* of the underlying utility,

$$\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < U^*_{im} \le \mu_{i1},\\
R_{im} &= 2 \text{ if } \mu_{i1} < U^*_{im} \le \mu_{i2},\\
R_{im} &= 3 \text{ if } \mu_{i2} < U^*_{im} \le \mu_{i3},\\
R_{im} &= 4 \text{ if } \mu_{i3} < U^*_{im} \le \mu_{i4},\\
R_{im} &= 5 \text{ if } \mu_{i4} < U^*_{im} < \infty.
\end{aligned} \tag{1.1}$$

The crucial feature of the description thus far is that the viewer has (and presumably knows) a continuous range of preferences that they could express if they were not forced to provide only an integer from one to five. Therefore, the observed rating represents a censored version of the true underlying preferences. Providing a rating of five could be an outcome ranging from general enjoyment to wild enthusiasm. Note that the *thresholds*, μ_{ij}, are specific to the person, and number (J-1) where J is the number of possible ratings (here, five) with $J - 1$ values needed to divide the range of utility into J cells. The thresholds are an important element of the model; they divide the range of utility into cells that are then identified with the observed ratings. One of the admittedly unrealistic assumptions in many applications is that these threshold values are the same for all individuals. Importantly, the difference between two levels of a rating scale (e.g., one compared to two, two compared to three) is not the same on a utility scale; hence we have a strictly nonlinear transformation captured by the thresholds, which are estimable parameters in an ordered choice model.

The model as suggested thus far provides a crude description of the mechanism underlying an observed rating. But it is simple to see how it might be improved. Any individual brings their own set of *characteristics* to the utility function, such as age, income, education, gender, where they live, family situation and so on, which we denote x_{i1}, $x_{i2}, \ldots, x_{iK}$. They also bring their own aggregate of unmeasured and unmeasurable (by the analyst) idiosyncrasies, denoted ε_{im}. How these features enter the utility function is uncertain, but it is conventional to use a linear function, which produces a familiar *random utility function*,

$$U_{im}^* = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{iK}x_{iK} + \varepsilon_{im}. \tag{1.2}$$

Once again, the model accommodates the intrinsic heterogeneity of individuals by allowing the coefficients to vary across them. To see how the heterogeneity across individuals might enter the ordered choice model, consider the user ratings of the same movie noted earlier, posted on December 1, 2008 at a different website, www.IMDb.com, as shown in Figure 1.1. This site uses a ten-point scale. The panel at the left below shows the overall ratings for 41,771 users of the site. The panel at the right shows how the average rating varies across age, gender and whether the rater is a US viewer or not.

An obvious shortcoming of the model is that otherwise similar viewers might naturally feel more enthusiastic about certain genres of movies (action, comedy, crime, etc.) or certain directors, actors or studios. It would be natural

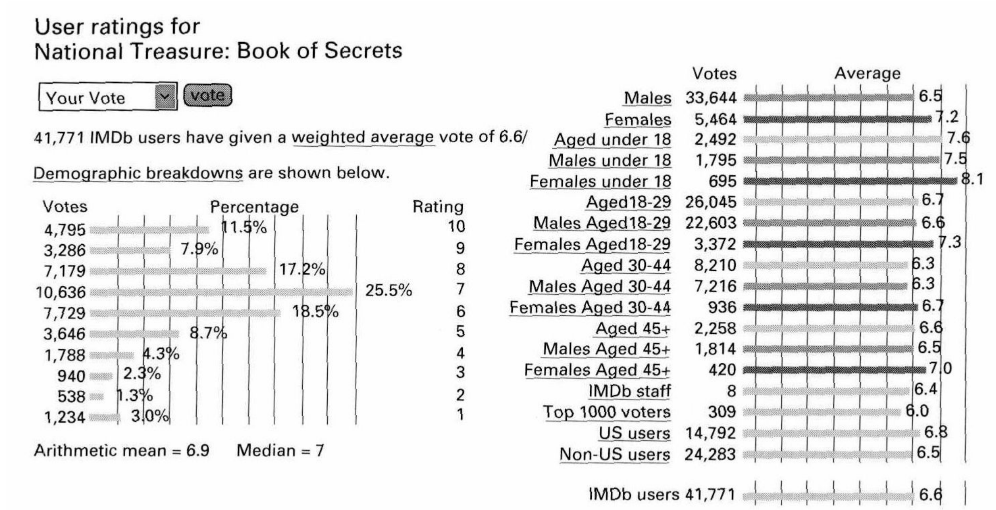

Figure 1.1 IMDb.com ratings (www.imdb.com/title/tt0465234/ratings)

for the utility function defined over movies to respond to certain *attributes* $z_1, z_2, \ldots, z_M$. The utility function might then appear, using a vector notation for the characteristics and attributes, as

$$U^*_{im} = \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im}. \tag{1.3}$$

Note, again, the marginal utilities of the attributes, $\boldsymbol{\delta}_i$, will vary from person to person. We note, finally, two possible refinements to accommodate additional sources of randomness, i.e., individual heterogeneity. Two otherwise *observably* identical individuals (same $\mathbf{x}_i$) seeing the same movie (same $\mathbf{z}_m$) might still react differently because of individual idiosyncrasies that are characteristics of the person that are the same for all movies. Some individuals are drawn to comedies and have low regard for dramas, while others might be uninterested in these two genres and enjoy only action movies. Second, every movie has unique features that are not captured by a simple *hedonic index* of its attributes – a particularly skillful character development, etc. A more complete *random utility function* might appear

$$U^*_{im} = \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m. \tag{1.4}$$

Finally, note that Netflix maintains a (huge) database of the ratings made by its users, including a complete history for each individual.

To return to the rating mechanism, the model we have constructed is:

$$
\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m \leq \mu_{i1},\\
R_{im} &= 2 \text{ if } \mu_{i1} < \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m \leq \mu_{i2},\\
R_{im} &= 3 \text{ if } \mu_{i2} < \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m \leq \mu_{i3},\\
R_{im} &= 4 \text{ if } \mu_{i3} < \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m \leq \mu_{i4},\\
R_{im} &= 5 \text{ if } \mu_{i4} < \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im} + u_i + v_m \leq \infty.
\end{aligned}
\tag{1.5}
$$

Perhaps relying on a central limit theorem to aggregate the innumerable small influences that add up to the individual idiosyncrasies and movie attraction, we assume that the random components, ε_{im}, u_i and v_m are normally distributed with zero means and (for now) constant variances. The assumption of normality will allow us to attach probabilities to the ratings. In particular, arguably the most interesting one is

$$
\text{Prob}(R_{im} = 5|\mathbf{x}_i, \mathbf{z}_m, u_i, v_m) = \text{Prob}[\varepsilon_{im} > \mu_{i4} - (\boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + u_i + v_m)].
\tag{1.6}
$$

The structure provides the framework for an econometric model of how individuals rate movies (that they rent from Netflix). The resemblance of this model to familiar models of binary choice is more than superficial. For example, one might translate this econometric model directly into a *probit model* by focusing on the variable

$$
\begin{aligned}
E_{im} &= 1 \text{ if } R_{im} = 5\\
E_{im} &= 0 \text{ if } R_{im} < 5.
\end{aligned}
\tag{1.7}
$$

Thus, our model is an extension of a binary choice model to a setting of more than two choices. However, the crucial feature of the model is the ordered nature of the observed outcomes and the correspondingly ordered nature of the underlying preference scale.

Beyond the usefulness of understanding the behavior of movie viewers, e.g., whether certain genres are more likely to receive high ratings or whether certain movies appeal to particular demographic groups, such a model has an additional utility to Netflix. Each time a subscriber logs on to the website after returning a movie, a computer program generates recommendations of other movies that it thinks that the viewer would enjoy (i.e., would give a rating of 5). The better the recommendation system is, the more attractive will be the website. Thus, the ability to predict accurately a "5" rating is a model feature that would have business value to Netflix. Netflix is currently (2008 until 2011) running a contest with a $1,000,000 prize to the individual who can devise the best algorithm for matching individual ratings based on ratings of other movies that they have rented. See www.netflixprize.com, Hafner (2006) and Thompson (2008). The Netflix prize and internet rating systems in general, beyond a large popular interest, have attracted a considerable amount of academic attention. See, for example, Ansari *et al.* (2000), Bennett and Lanning (2007) and Umyarov and Tuzhlin (2008).

The model described here is an *ordered choice model.* (The choice of the normal distribution for the random term makes it an *ordered probit model.*) Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential ingredient is the *mapping from an underlying, naturally ordered preference scale to a discrete, ordered observed outcome,* such as the rating scheme described above. The model of ordered choice pioneered by Aitchison and Silvey (1957) and Snell (1964) and articulated in its modern form by Zavoina and McKelvey (1975), McKelvey and Zavoina (1971, 1975), and McCullagh (1980) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly. A search of just the "ordered probit" model identified applications on:

- academic grades (Butler *et al.* (1994), Li and Tobias (2006a));
- bond ratings (Terza (1985));
- Congressional voting on a Medicare bill (McKelvey and Zavoina (1975));
- credit ratings (Cheung (1996), Metz and Cantor (2006));
- driver injury severity in car accidents (Eluru *et al.* (2008), Wang and Kockelman (2008));
- drug reactions (Fu *et al.* (2004));
- duration (Han and Hausman (1986, 1990), Ridder (1990));
- education (Carneiro *et al.* (2001, 2003), Machin and Vignoles (2005), Cameron and Heckman (1998), Johnson and Albert (1999), Cunha *et al.* (2007));
- eye disease severity (Biswas and Das (2002));
- financial failure of firms (Jones and Hensher (2004), Hensher and Jones (2007));
- happiness (Winkelmann (2005), Zigante (2007));
- health status (Riphahn *et al.* (2003), Greene (2008a));
- insect resistance to insecticide (Walker and Duncan (1967));
- job classification in the military (Marcus and Greene (1983));
- job training (Groot and van den Brink (2003c));