

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

	<i>List of tables</i>	page ix
	<i>List of figures</i>	xi
	<i>Preface</i>	xiii
1	Introduction: random utility and ordered choice models	1
2	Modeling binary choices	9
	2.1 Random utility formulation of a model for binary choice	10
	2.2 Probability models for binary choices	11
	2.3 Estimation and inference	16
	2.4 Covariance matrix estimation	26
	2.5 Application of the binary choice model to health satisfaction	28
	2.6 Partial effects in a binary choice model	30
	2.7 Hypothesis testing	39
	2.8 Goodness of fit measures	44
	2.9 Heteroscedasticity	54
	2.10 Panel data	57
	2.11 Parameter heterogeneity	75
	2.12 Endogeneity of a right-hand-side variable	80
	2.13 Bivariate binary choice models	83
	2.14 The multivariate probit and panel probit models	93
	2.15 Endogenous sampling and case control studies	96
3	A model for ordered choices	99
	3.1 A latent regression model for a continuous measure	99
	3.2 Ordered choice as an outcome of utility maximization	103
	3.3 An observed discrete outcome	105
	3.4 Probabilities and the log likelihood	108

3.5	Log likelihood function	108
3.6	Analysis of data on ordered choices	109
4	Antecedents and contemporary counterparts	111
4.1	The origin of probit analysis: Bliss (1934a), Finney (1947a)	111
4.2	Social science data and regression analysis for binary outcomes	116
4.3	Analysis of binary choice	117
4.4	Ordered outcomes: Aitchison and Silvey (1957), Snell (1964)	117
4.5	Minimum chi squared estimation of an ordered response model: Gurland <i>et al.</i> (1960)	123
4.6	Individual data and polychotomous outcomes: Walker and Duncan (1967)	125
4.7	McKelvey and Zavoina (1975)	127
4.8	Cumulative odds model	127
4.9	Continuation ratio model	130
4.10	The ordered regression model	130
4.11	Other related models	132
4.12	The latent continuous variable	134
5	Estimation, inference and analysis using the ordered choice model	136
5.1	Application of the ordered choice model to self-assessed health status	136
5.2	Distributional assumptions	138
5.3	The estimated ordered probit (logit) model	138
5.4	The estimated threshold parameters	140
5.5	Interpretation of the model – partial effects and scaled coefficients	142
5.6	Inference	150
5.7	Prediction – computing probabilities	157
5.8	Measuring fit	160
5.9	Estimation issues	167
6	Specification issues and generalized models	181
6.1	Functional form issues and the generalized ordered choice model (1)	181
6.2	Model implications for partial effects	193
6.3	Methodological issues	198
6.4	Specification tests for ordered choice models	198

7	Accommodating individual heterogeneity	208
7.1	Threshold models – the generalized ordered probit model (2)	209
7.2	Nonlinear specifications – a hierarchical ordered probit (HOPIT) model	214
7.3	Thresholds and heterogeneity – anchoring vignettes	219
7.4	Heterogeneous scaling (heteroscedasticity) of random utility	232
7.5	Individually heterogeneous marginal utilities	237
	Appendix: Equivalence of the vignette and HOPIT models	237
8	Parameter variation and a generalized model	239
8.1	Random-parameters models	239
8.2	Latent class and finite mixture modeling	247
8.3	Generalized ordered choice model with random thresholds (3)	262
9	Ordered choice modeling with panel and time series data	268
9.1	Ordered choice models with fixed effects	268
9.2	Ordered choice models with random effects	275
9.3	Testing for random or fixed effects: a variable addition test	278
9.4	Extending parameter heterogeneity models to ordered choices	281
9.5	Dynamic models	285
9.6	Spatial autocorrelation	289
10	Bivariate and multivariate ordered choice models	290
10.1	Multiple equations	290
10.2	Bivariate ordered probit models	291
10.3	Polychoric correlation	294
10.4	Semi-ordered bivariate probit model	295
10.5	Applications of the bivariate ordered probit model	295
10.6	A panel data version of the bivariate ordered probit model	297
10.7	Trivariate and multivariate ordered probit models	299
11	Two-part and sample selection models	302
11.1	Inflation models	302
11.2	Sample selection models	306
11.3	An ordered probit model with endogenous treatment effects	319
12	Semiparametric and nonparametric estimators and analyses	320
12.1	Heteroscedasticity	321
12.2	A distribution free estimator with unknown heteroscedasticity	323

12.3	A semi-nonparametric approach	324
12.4	A partially linear model	327
12.5	Semiparametric analysis	327
12.6	A nonparametric duration model	329
<i>References</i>		337
<i>Index</i>		361

Tables

2.1	Data used in binary choice application	<i>page</i> 29
2.2	Estimated probit and logit models	29
2.3	Alternative estimated standard errors for the probit model	30
2.4	Partial effects for probit and logit models at means of x	32
2.5	Marginal effects and average partial effects	37
2.6	Hypothesis tests	43
2.7	Homogeneity test	43
2.8	Fit measures for probit model	49
2.9a	Prediction success for probit model based on $\hat{y}_i = 1[\hat{F}_i > .5]$	50
2.9b	Predictions for probit model based on probabilities	50
2.10	Success measures for predictions by estimated probit model using $\hat{y}_i = 1[\hat{F}_i > .5]$	50
2.11	Heteroscedastic probit model	56
2.12	Cluster corrected covariance matrix (7,293 groups)	59
2.13	Fixed effects probit model	61
2.14	Estimated fixed effects logit models	65
2.15	Estimated random effects probit models	70
2.16a	Semiparametric random effects probit model	70
2.16b	Estimated parameters for four class latent class model	70
2.17	Random effects model with Mundlak correction	74
2.18	Estimated random parameter models	78
2.19	Estimated partial effects	78
2.20	Cross-tabulation of healthy and working	86
2.21	Estimated bivariate probit model	87
2.22	Estimated sample selection model	93
2.23	Estimated panel probit model	95
4.1	McCullagh application of an ordered outcome model	130
5.1	Estimated ordered choice models: probit and logit	139
5.2	Estimated partial effects for ordered choice models	144
5.3	Estimated expanded ordered probit model	147
5.4	Transformed latent regression coefficients	150

5.5	Estimated partial effects with asymptotic standard errors	158
5.6	Mean predicted probabilities by kids	159
5.7a	Predicted vs. actual outcomes for ordered probit model	165
5.7b	Predicted probabilities vs. actual outcomes for ordered probit model	165
5.8	Predicted vs. actual outcomes for automobile data	166
5.9	Grouped data for ordered choice modeling response frequency in a taste-testing experiment	168
5.10	Estimated ordered choice model based on grouped data	169
5.11	<i>Stata</i> and <i>NLOGIT</i> estimates of an ordered probit model	172
5.12	Software used for ordered choice modeling	179
6.1	Brant test for parameter homogeneity	186
6.2	Estimated ordered logit and generalized ordered logit (1)	191
6.3	Boes and Winkelmann estimated partial effects	194
7.1	Estimated generalized ordered probit models from Terza (1985)	211
7.2	Estimated hierarchical ordered probit models	217
7.3	Estimated partial effects for ordered probit models	218
7.4	Predicted outcomes from ordered probit models	219
7.5	Estimated heteroscedastic ordered probit model	235
7.6	Partial effects in heteroscedastic ordered probit model	236
8.1	Estimated random parameters ordered probit model	242
8.2	Implied estimates of parameter matrices	243
8.3	Estimated partial effects from random-parameters model	244
8.4	Estimated two-class latent class ordered probit models	256
8.5	Estimated partial effects from latent class models	257
8.6	Estimated generalized random-thresholds ordered logit model	266
9.1	Monte Carlo analysis of the bias of the MLE in fixed-effects discrete choice models (Means of empirical sampling distributions, $n = 1,000$ individuals, $R = 200$ replications)	270
9.2	Fixed-effects ordered choice models	279
9.3	Random effects ordered logit models – quadrature and simulation	280
9.4	Random effects ordered probit model with Mundlak correction	281
9.5	Random parameters ordered logit model	283
9.6	Latent-class ordered logit models	284
10.1	Applications of bivariate ordered probit since 2000	296
11.1	Estimated ordered probit sample selection model	310
12.1	Grouping of strike durations	334
12.2	Estimated logistic duration models for strike duration	334

Figures

1.1	IMDb.com ratings (www.imdb.com/title/tt0465234/ratings)	page 3
2.1	Random utility basis for a binary outcome	11
2.2	Probability model for binary choice	14
2.3	Probit model for binary choice	15
2.4	Partial effects in a binary choice model	31
2.5	Fitted probabilities for a probit model	39
2.6	Prediction success for different prediction rules	51
2.7	ROC curve for estimated probit model	54
2.8	Distribution of conditional means of income parameter	80
3.1	Underlying probabilities for an ordered choice model	108
4.1	Insecticide experiment	112
4.2	Table of probits for values of p_i	114
4.3	Percentage errors in Pearson table of probability integrals	114
4.4	Implied spline regression in Bliss's probit model	115
5.1	Self-reported health satisfaction	137
5.2	Health satisfaction with combined categories	137
5.3	Estimated ordered probit model	140
5.4a	Sample proportions	141
5.4b	Implied partitioning of latent normal distribution	141
5.5	Partial effect in ordered probit model	145
5.6	Predicted probabilities for different ages	160
6.1	Estimated partial effects in Boes and Winkelmann (2006b) models	196
6.2	Estimated partial effects for linear and nonlinear index functions	196
7.1	Differential item functioning in ordered choices	220
7.2	KMST comparison of political efficacy	229
7.3	KMST estimated vignette model	230
8.1	Kernel density for estimate of the distribution of means of income coefficient	247

12.1	Table 1 from Stewart (2005)	325
12.2	Job satisfaction application, extended	326
12.3	Strike duration data	333
12.4	Estimated nonparametric hazard functions	335
12.5	Estimated hazard function from log-logistic parametric model	335

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,

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

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 $x_{i1}, x_{i2}, \dots, x_{iK}$. 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*,

$$U_{im}^* = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \dots + \beta_{iK}x_{iK} + \varepsilon_{im}. \quad (1.2)$$

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

User ratings for
National Treasure: Book of Secrets

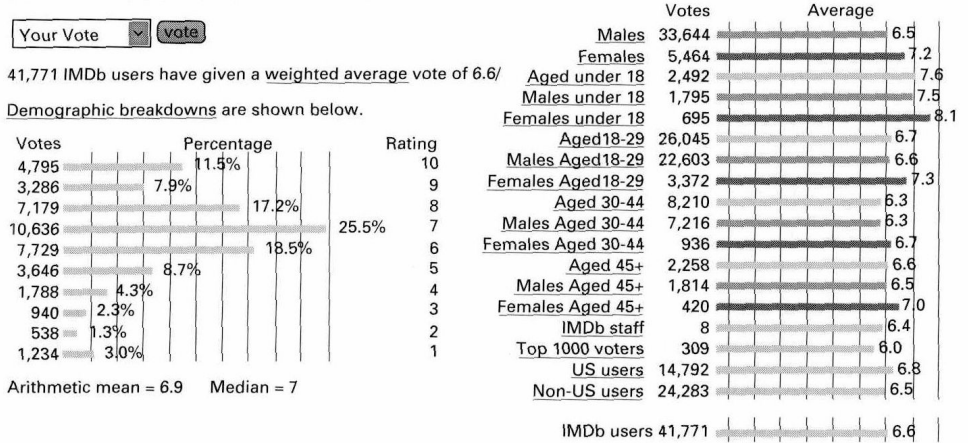


Figure 1.1 IMDb.com ratings (www.imdb.com/title/tt0465234/ratings)

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

$$U_{im}^* = \beta_i' \mathbf{x}_i + \delta_i' \mathbf{z}_m + \varepsilon_{im}. \quad (1.3)$$

Note, again, the marginal utilities of the attributes, δ_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

$$U_{im}^* = \beta_i' \mathbf{x}_i + \delta_i' \mathbf{z}_m + \varepsilon_{im} + u_i + v_m. \quad (1.4)$$

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:

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

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

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

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), Zigarette (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));