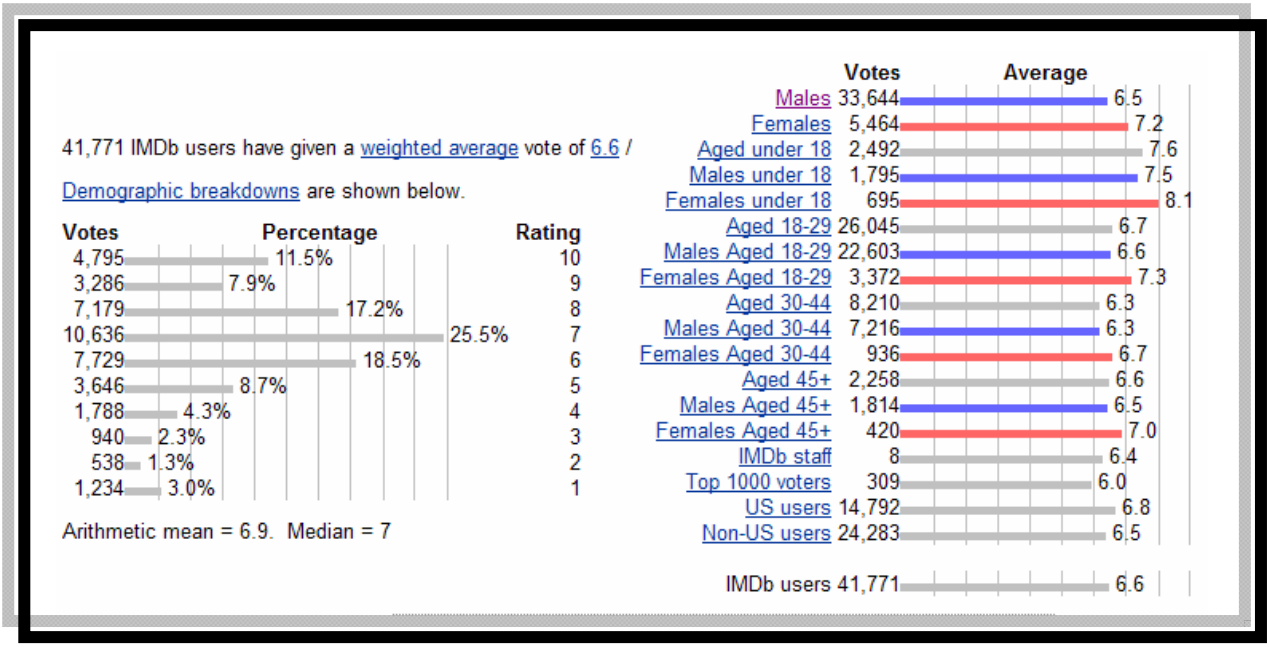
Modeling Ordered Choices



William H. Greene[1]

David A. Hensher[2]

January, 2009

[1]Department of Economics, Stern School of Business, New York University, New York, NY 10012, wgreene@stern.nyu.edu

[2]Institute of Transport and Logistics Studies, Faculty of Economics and Business, University of Sydney, NSW 2006 Australia Hensher@itls.usyd.edu.au

Brief Contents

Contents

List of Tables

List of Figures

Preface

This book began as a short note to propose the new estimator in Section 8.3. In researching the recent developments in ordered choice modeling, we decided 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 using grouped data. But, these stop well short of examining the extensive selection 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. This motivated us to assemble this more complete overview of the topic.

We strongly believe that many practitioners (and theorists) focus too sharply on coefficient estimation and do not place enough attention on the meaning of the model or its components. As this review proceeded, it struck us that a more thorough survey of the model, itself, including its historical development might be useful and (we hope) interesting for readers. The following is also a survey of the methodological literature on the model of ordered choice. (We have, of necessity, omitted mention of many – perhaps most – of the huge number of applications.)

The development of the ordered choice regression model has emerged in two surprisingly disjoint strands of literature, in its earliest forms in the bioassay literature and in its modern social science counterpart with the pioneering paper by McElvey 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 bibliography of this survey shares only four items with its 100+ entry counterpart in Johnson and Albert (*Ordinal Data Modeling*, 1999).]

The earliest applications of modeling ordered outcomes involved aggregate 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 40 covariates – and many levels of those inputs, then crosstabulations 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. McElvey and Zavoina (1975) added to the model a reliance on a formal underlying "data generating process," the latent regression, a mechanism that 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 an introduction to a certain class of discrete choice models. We anticipate that it can be used in a graduate level course in econometrics or statistics after the first one at the level of, say, Greene (2008a) and as a reference in specialized courses such as microeconometrics or discrete choice modeling. The range of applications of ordered choice models 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 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. We leave it to others, e.g., Wooldridge (2002a), Hayashi (2000) 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.) They can also be done with varying degrees of difficulty with several other packages, such as Stata and SAS. Since this book is not a 'how to' for any particular computer program, we have not provided any instruction 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 program, 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.

William H. Greene
David A. Hensher
New York, January, 2009

1

Introduction: Random Utility 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 5 is the highest, most favorable rating. The ratings of the many thousands of subscribers who rented that movie are

Figure 1.1 Netflix Film Average Rating

averaged to provide a recommendation to prospective viewers, as shown for example in Figure 1.1. 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 having the following range:

$$-\infty < U_{im}^* < +-\infty$$

where i indicates the individual and m indicates the movie. Individuals are invited to "rate" the movie on an integer scale from 1 to 5. 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 5 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) – J-1 values are 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., 1 compared to 2, 2 compared to 3) 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},…,x_{iK}. They also bring their own aggregate of unmeasured and unmeasurable (by the statistician) idiosyncracies, 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*,

$$U_{im}^* = \beta_{i0} + \beta_{i1}\mathbf{x}_{i1} + \beta_{i2}\mathbf{x}_{i2} + \ldots + \beta_{iK}\mathbf{x}_{iK} + \varepsilon_{im}. \qquad (1.2)$$

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

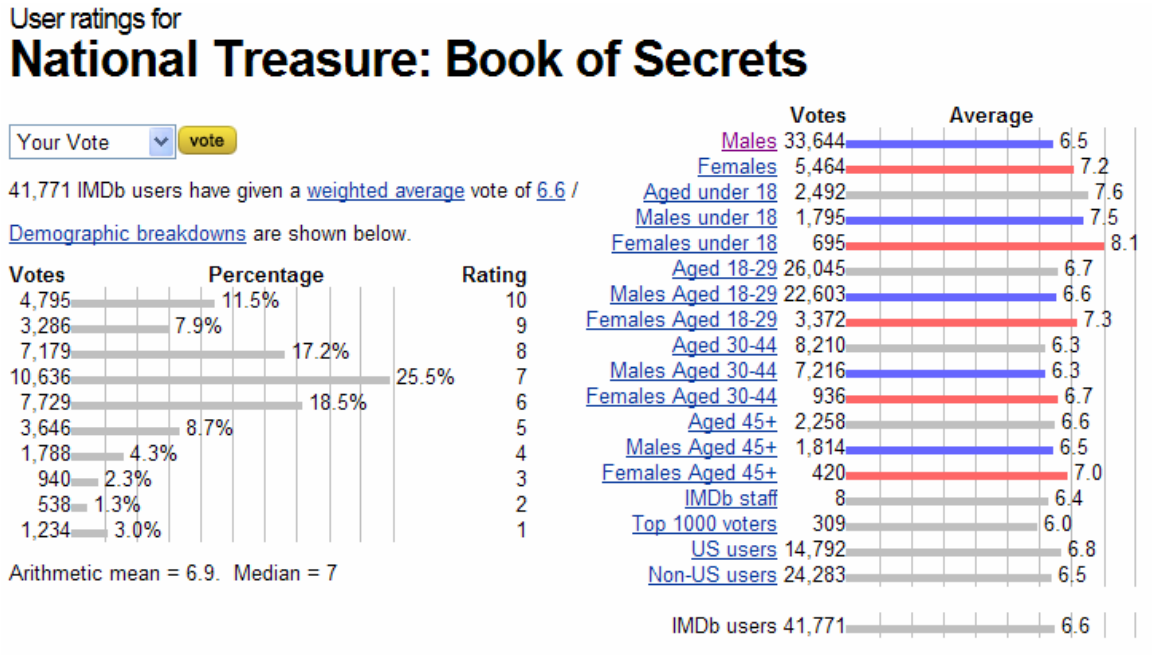


Figure 1.2 IMDB.com Ratings (http://www.imdb.com/title/tt0465234/ratings)

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

$$U_{im}^* = \boldsymbol{\beta}_i'\mathbf{x}_i + \boldsymbol{\delta}_i'\mathbf{z}_m + \varepsilon_{im}. \qquad (1.3)$$

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 (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 idiosyncracies that are characteristics of the person that are the same for all 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 relatively complete utility function might appear

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

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

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

Perhaps relying on a central limit to aggregate the innumerable small influences that add up to the individual idiosyncracies 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

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

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, we see the model is an extension of a binary choice model to a setting of more than two choices. But, we emphasize, 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 accurately to predict a "5" rating is a model feature that would have business value to Netfix. 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 Thomson (2008). The Netflix prize and internet rating systems in general, beyond a large popular interest, have attracted a considerable amount of academic attention as well. See, for example,

Ahsari, Essegaier and Kohli (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 some underlying, naturally ordered preference scale to an ordered observed outcome*, such as the rating scheme described above. The model of ordered choice pioneered by Aitcheson and Silvey (1957) and Snell (1964) and articulated in its modern form by McElvey and Zavoina (1969, 1971, 1975) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly. A quick 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 [McElvey and Zavoina (1975)],
- credit ratings [Cheung (1996)],
- driver injury severity in car accidents [Eluru, Bhat and Hensher (2008)],
- drug reactions [Fu et al.(2004)],
- duration [Han and Hausman (1990), Ridder (1990)],
- education [Machin and Vignoles (2005), Carneiro, Hansen and Heckman (2001, 2003), Cameron and Heckman (1998), Cunha, Heckman and Navarro (2007), Johnson and Albert (1999)],
- 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 [Greene (2008a) based on Riphahn, Wambach and Million (2003)],
- insect resistance to insecticide [Walker and Duncan (1967)],
- job classification in the military [Marcus and Greene (1983)],
- job training [Groot and van den Brink (2002a)],
- labor supply [Heckman and MaCurdy (1981)],
- life satisfaction [Clark et al. (2001), Wim and ven den Brink (2002, 2003b)],
- monetary policy [Eichengreen, Watson and Grossman (1985)],
- nursing labor supply [Brewer et al. (2008)],
- obesity [Greene, Harris, Hollingsworth and Maitra (2008)],
- perceptions of difficulty making left turns [Zhang (2007)],
- pet ownership [Butler and Chatterjee(1997)],
- political efficacy (a cross country comparison) [King et al. (2004)],
- product quality [Prescott and Visscher (1977), Shaked and Sutton (1982)],
- promotion and rank in nursing [Pudney and Shields (2000)],
- stock price movements [Tsay (2005)],
- tobacco use [Harris and Zhao (2007), Kasteridis, Munkin and Yen (2008)],
- trip stops [Bhat (1997)],
- vehicle ownership [Bhat and Pulugurta (1998), Train (1986), Hensher, Smith, Milthorpe and Bernard (1992),
- work disability [Kapteyn et al. (2007)]

and hundreds more.

This book will survey the development and use of models of ordered choices from the perspective of the social sciences. The distinction between that and the biological sciences will

emerge clearly as we proceed. We will detail the model itself, estimation and inference, interpretation and analysis. We will also survey a wide variety of different kinds of applications, and a wide range of variations and extensions on the basic model that have been proposed in the recent literature.

The practitioner who desires a quick entry level primer on the model can choose among numerous sources for a satisfactory introduction to the ordered choice model and its uses. Social science oriented introductions to the ordered choice model appear in journal articles such as Winship and Mare (1984), Becker and Kennedy (1992), Daykin and Moffatt (2002) and Boes and Winkelmann (2006a), and in textbook and monograph treatments including Maddala (1983), DeMaris (2004), Long (1997), Johnson and Albert (1999), Long and Freese (2006) and Greene (2008a). There are also many surveys and primers for bioassay, including, e.g., Greenland (1994), Agresti (1999) and Ananth and Kleinbaum (1997). This survey is offered as an addition to this list largely to broaden the discussion of the model and for a number of specific purposes:

- Many interesting extensions of the model already appearing in the literature are not mentioned in the surveys listed above.
- Recent analyses of the ordered choice model have uncovered some interesting avenues of generalization.
- The model formulation rests on a number of subtle underlying aspects that are not developed as completely as are the mechanics of using the "technique." Only a few of the surveys devote substantial space to interpreting the model's components once they are estimated. As made clear here and elsewhere, the coefficients in an ordered choice provide, in isolation, provide little useful information about the phenomenon under study. Yet, estimation of coefficients and tests of statistical significance are the central (sometimes, only) issue in many of the surveys listed above, and in some of the received applications.
- We will offer our own generalizations of the ordered choice model.
- With the creative development of easy to use contemporary software, many model features and devices are served up because they *can* be computed without much (or any) discussion of *why* they would be computed, or, in some cases, even *how* they are computed. To cite an example, Long and Freese (2006, pp. 195-196) state "several different measures [of fit] can be computed…" [using Stata] for the ordered probit model. Their table that follows lists 20 values, seven of which are statistics whose name contains "R squared." The values range from 0.047 to 0.432. No discussion of what the measures are, what they mean, or how they are computed follows; the section provides the reader with a single statement that two Monte Carlo studies have found that one of the measures "closely approximates the R^2 obtained by fitting the linear regression model on the underlying latent variable." (Note that the underlying variable – utility in our earlier example – is never observed.) Obviously researchers differ on what information they wish to extract from the data. We will attempt to draw the focus to a manageable few aspects of the model that appear to have attained some degree of consensus.

The book proceeds as follows. Standard models of binary choice are presented in Chapter 2. The fundamental ordered choice model is developed in some detail in Chapter 3. The historical antecedents to the basic ordered choice model are documented in Chapter 4. In Chapter 5, we return to the modern form of the model, and develop the different aspects of its use, such as interpreting the model, statistical inference and fit measures. Some recent generalizations and extensions are presented in Chapters 6 - 11. Semiparametric models that reach beyond the

mainstream of research are discussed in Chapter 12. An application based on a recent study of health care [Riphahn, Wambach and Million (2003)] will be dispersed through the discussion to provide an illustration of the points being presented.

There is a large literature parallel to the social science applications in the areas of biometrics and psychometrics. The distinction is not perfectly neat, but there is a tangible difference in orientation, as will be evident below. From the beginning with Bliss's (1934a) invention of probit modeling, many of the methodological and statistical developments in the area of ordered choice modeling have taken place in this setting. It will be equally evident that these two areas of application have developed in parallel, but by no means in concert. This book is largely directed toward social science applications. However, the extensions and related features of the models and techniques in biometrics will be integrated into the presentation.

2

Modeling Binary Choices

The *random utility* model described in the Introduction is one of two essential building blocks that form the foundation for modeling ordered choices. The second fundamental pillar is the *model for binary choices*. The ordered choice model that will be the focus of the rest of this book is an extension of a model used to analyze the situation of a choice between two alternatives – whether the individual takes an action or does not, or chooses one of two elemental alternatives, and so on. This chapter will develop the standard model for binary choices in considerable detail. Many of the results analyzed in the chapters to follow will then be straightforward extensions.

We present a lengthy survey of binary choice modeling. There are numerous such surveys available, including Amemiya (1981), Greene (2008a, Chapter 23) and several book length treatments such as Cox (1970). Our interest here is in the aspects of binary choice modeling that are likely to reappear in the analysis of ordered choices. We have therefore bypassed numerous topics that do appear in other treatments, notably semiparametric and nonparametric approaches, but whose counterparts have not yet made significant inroads in ordered choice modeling. (Chapter 12 does contain some description of a few early entrants to this nascent literature.) This chapter also contains a long list of topics related to binary choice modeling, such as fit measures, multiple equation models, sample selection and many others, that are useful as components or building blocks in the analysis of ordered choices. Our intent with this chapter is to extend beyond conventional binary choice modeling, and provide a bridge to the somewhat more involved models for ordered choices. Quite a few of these models, such as the sample selection model, are straightforward to generalize to the ordered probit model.

The orientation of our treatment is the analysis of individual choice data, as typically appears in social science applications using survey data. An example is the application developed below in which survey data on health satisfaction are transformed into a binary outcome that states whether or not a respondent feels healthier than average. A parallel literature in, e.g., bioassay such as Cox (1970) and Johnson and Albert (1999) is often focused on 'grouped' data in the form of proportions. Two examples would be an experiment to determine the lethality of a new insecticide in which n_i insects are subjected to dosage x_i, and a proportion p_i succumb to the dose, and a state by state tally of voting proportions in a national election. With only a few exceptions noted in passing, we will not be concerned with data of this type.

2.1 Random Utility Formulation of a Model for Binary Choice

An application we will develop is based on a survey question in a large German panel data set, roughly, "on a scale from zero to ten, how satisfied are you with your health?" The full data set consists of from one to seven observations – it is an unbalanced panel – on 7,293 households for a total of 27,326 family year observations. A histogram of the responses appears in Figure 5.1. Consistent with the description in the Introduction, we might formulate a random utility/ordered choice model for the variable R_i = "*Health Satisfaction*" as

$$
\begin{aligned}
U_i^* &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \\
R_i &= 0 \text{ if } -\infty < U_i^* \leq \mu_0, \\
R_i &= 1 \text{ if } \mu_0 < U_i^* \leq \mu_1, \\
&\cdots \\
R_i &= 10 \text{ if } \mu_9 < U_i^* < +\infty,
\end{aligned}
$$

where $\mathbf{x}_i$ is a set of variables such as gender, income, age, and education that are thought to influence the response to the survey question. (Note that at this point, we are pooling the panel data as if it were a cross section of n = 32,726 independent observations and denoting by i one of those observations.) The average response in the full sample is a bit less than 7. Consider a simple response variable, y_i = "*Healthy*,"

$$y_i = 1 \text{ if } R_i \geq 7 \text{ and } y_i = 0 \text{ otherwise.}$$

Then, in terms of the original variables, the model for y_i is

$$y_i = 0 \text{ if } R_i = 0, 1, 2, 3, 4, 5 \text{ or } 6.$$

By adding the terms, we then find, for the two possible outcomes,

$$y_i = 0 \text{ if } U_i^* \leq \mu_6,$$
$$y_i = 1 \text{ if } U_i^* > \mu_6.$$

Figure 2.1 shows how the variable y_i is generated from the underlying utility.

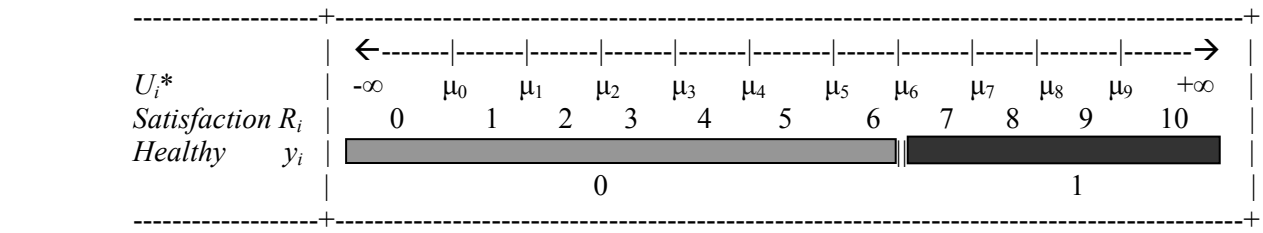


Figure 2.1 Random Utility Basis for a Binary Outcome

Substituting for U_i*, we find

$$y_i = 1 \text{ if } \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i > \mu_6$$
or $$y_i = 1 \text{ if } \varepsilon_i > \mu_6 - \boldsymbol{\beta}'\mathbf{x}_i$$
and $$y_i = 0 \text{ otherwise.}$$

We now assume that the first element of $\boldsymbol{\beta}'\mathbf{x}_i$ is a constant term, α, so that $\boldsymbol{\beta}'\mathbf{x}_i - \mu_6$ equivalent to $\boldsymbol{\gamma}'\mathbf{x}_i$ where the first element of $\boldsymbol{\gamma}$ is $\alpha - \mu_6$ and the rest of $\boldsymbol{\gamma}$ is the same as the rest of $\boldsymbol{\beta}$. Then, the binary outcome is determined by

$$y_i = 1 \text{ if } \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i > 0$$
and $$y_i = 0 \text{ otherwise.}$$

In general terms, we write our binary choice model in terms of the underlying utility as

$$y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i,$$
$$y_i = 1[y_i^* > 0], \tag{2.1}$$

where the function 1[condition] equals one if the condition is true and zero if it is false.

2.2 Probability Models for Binary Choices

The observed outcome, y_i, is determined by a *latent regression*,

$$y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i.$$

The random variable y_i takes two values, one and zero, with probabilities

$$\begin{aligned} \text{Prob}(y_i = 1|\mathbf{x}_i) &= \text{Prob}(y_i^* > 0|\mathbf{x}_i) \\ &= \text{Prob}(\boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i > 0) \\ &= \text{Prob}(\varepsilon_i > -\boldsymbol{\gamma}'\mathbf{x}_i). \end{aligned} \tag{2.2}$$

The model is completed by the specification of a particular probability distribution for ε_i. In terms of building an internally consistent model, we require that the probabilities be between zero and one and that they increase when $\boldsymbol{\gamma}'\mathbf{x}_i$ increases. In principle, any probability distribution defined over the entire real line will suffice, though empirically, one might be interested in investigating whether one specification is preferable to another.

2.2.1 Nonparametric and Semiparametric Specifications

The fully parametric probit and logit models discussed in the rest of this chapter remain by far the mainstays of empirical research on binary choice. Fully nonparametric discrete choice models are fairly exotic and have made only limited inroads in the applied literature – though they have attracted a considerable attention in the more theoretical literature, e.g., Matzkin (1993). The primary obstacle to application is their paucity of interpretable results. [See Manski (1987, 1995).] Semiparametric estimators represent a compromise between the robust but thinly informative nonparametric estimators and fragile fully parametric approaches. Klein and Spady's (1993) model has been used in several applications, including Gerfin (1996), Horowitz (1993), and Fernandez and Rodriguez-Poo (1997). The single index formulation departs from a linear "regression" formulation,

$$E[y_i \mid \mathbf{x}_i] = E[y_i \mid \boldsymbol{\gamma}'\mathbf{x}_i].$$

Then

$$\text{Prob}(y_i = 1 \mid \mathbf{x}_i) = F(\boldsymbol{\gamma}'\mathbf{x}_i \mid \mathbf{x}_i) = G(\boldsymbol{\gamma}'\mathbf{x}_i),$$

where G is an unknown continuous distribution function whose range is [0, 1]. The function G is not specified a priori; it is estimated along with the parameters. The estimator of the probability function, G_n, is computed using a nonparametric kernel estimator of the density of $\boldsymbol{\gamma}'\mathbf{x}_i$. There is a large and burgeoning literature on kernel estimation and nonparametric estimation in econometrics. [An application is Melenberg and van Soest (1996).] Li and Racine (2007) is a comprehensive introduction to the subject.

2.2.2 The Linear Probability Model

The binary choice model is sometimes based on a *linear probability model* (LPM),

$$\text{Prob}(y_i = 1|\mathbf{x}_i) = \boldsymbol{\gamma}'\mathbf{x}_i.$$

The model has a fundamental flaw in that probabilities must lie between zero and one, but the linear function cannot be so constrained. Further discussion of the LPM may be found in Aldrich

and Nelson (1984), Amemiya (1981), Maddala (1983, Section 2.2) and Greene (2008a). Notwithstanding its shortcomings, the model has been employed in numerous applications, such as Caudill (1988), Heckman and MaCurdy (1985), Heckman and Snyder (1997) and Angrist (2001). Since the LPM has not played a role in the evolution of the ordered choice models, we will not consider it further.

2.2.3 The Probit and Logit Models

The literature is overwhelmingly dominated by two models, the standard normal distribution, which gives rise to the *probit model*,

$$f(\varepsilon_i) = \frac{\exp\left(-\varepsilon_i^2 / 2\right)}{\sqrt{2\pi}}, \tag{2.3}$$

and the standard logistic distribution, which produces the *logit model*. The logistic distribution,

$$f(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{[1+\exp(\varepsilon_i)]^2}, \tag{2.4}$$

resembles the normal distribution, but has somewhat thicker tails – it more closely resembles the *t* distribution with seven degrees of freedom. Other distributions, such as the complementary log log and Gompertz distribution that are built into modern software such as *Stata* and *NLOGIT* are sometimes specified as well, without obvious motivation. The normal distribution can be motivated by an appeal to the central limit theorem and modeling human behavior as the sum of myriad underlying influences. The logistic distribution has proved to be a useful functional form for modeling purposes for several decades. These two are by far the most frequently used in applications.

Figure 2.2 shows how the distribution of the underlying utility is translated into the probabilities for the binary outcomes for y_i. The shaded area is Prob(y_i = 1|$\mathbf{x}_i$) = Prob(ε_i > -$\boldsymbol{\gamma}'\mathbf{x}_i$).

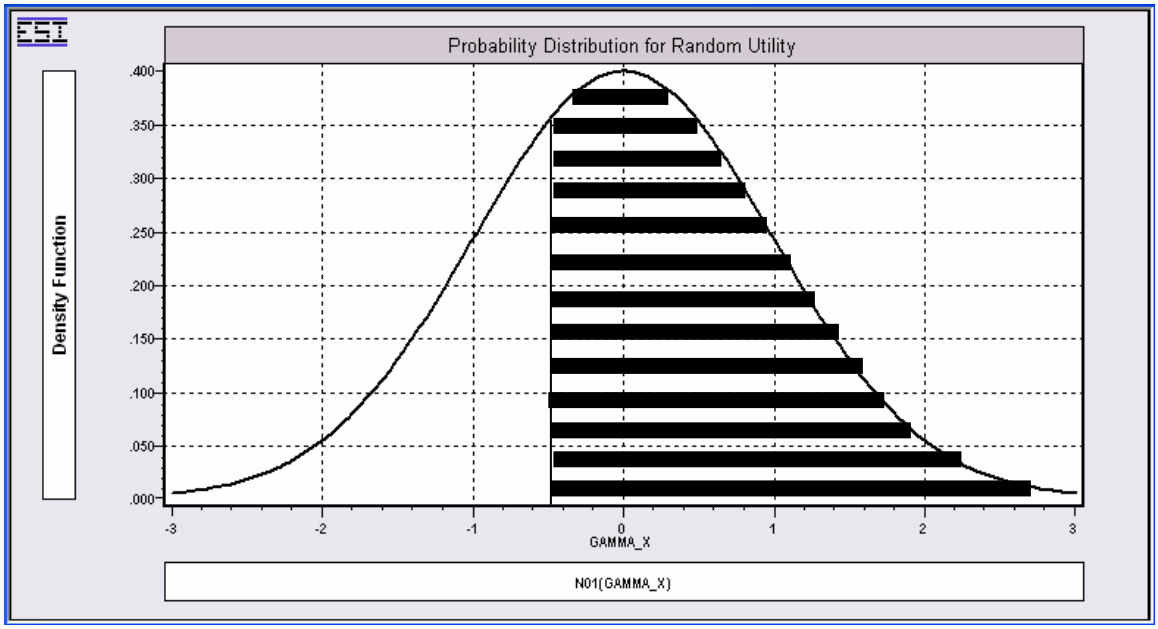


Figure 2.2. Probability Model for Binary Choice

The implication of the model specification is that $y_i|\mathbf{x}_i$ is a Bernoulli random variable with

$$\begin{aligned}\text{Prob}(y_i = 1|\mathbf{x}_i) &= \text{Prob}(y_i^* > 0|\mathbf{x}_i) \\ &= \text{Prob}(\varepsilon_i > -\boldsymbol{\gamma}'\mathbf{x}_i) \\ &= \int_{-\boldsymbol{\gamma}'\mathbf{x}_i}^{\infty} f(\varepsilon_i)d\varepsilon_i \\ &= 1 - F(-\boldsymbol{\gamma}'\mathbf{x}_i),\end{aligned} \tag{2.5}$$

where $F(.)$ denotes the cumulative density function (CDF) for ε_i. The standard normal and standard logistic distributions are both *symmetric distributions* that have the property that

$$F(\boldsymbol{\gamma}'\mathbf{x}_i) = 1 - F(-\boldsymbol{\gamma}'\mathbf{x}_i).$$

This produces the convenient result

$$\text{Prob}(y_i = 1|\mathbf{x}_i) = F(\boldsymbol{\gamma}'\mathbf{x}_i). \tag{2.6}$$

Standard notations for the normal and logistic distribution functions are

$$\begin{aligned}&\text{Prob}(y_i = 1|\mathbf{x}_i) = \Phi(\boldsymbol{\gamma}'\mathbf{x}_i) \text{ if } \varepsilon_i \text{ is normally distributed and} \\ &\text{Prob}(y_i = 1|\mathbf{x}_i) = \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i) \text{ if } \varepsilon_i \text{ is logistically distributed.}\end{aligned} \tag{2.7}$$

There is no closed form for the normal cdf, $\Phi(t)$; it is computed by approximation (usually by a ratio of polynomials.) But, the logistic cdf does exist in closed form,

$$\Lambda(t) = \exp(t) / [1 + \exp(t)].$$

The resulting probit model for a binary outcome is shown in Figure 2.3.

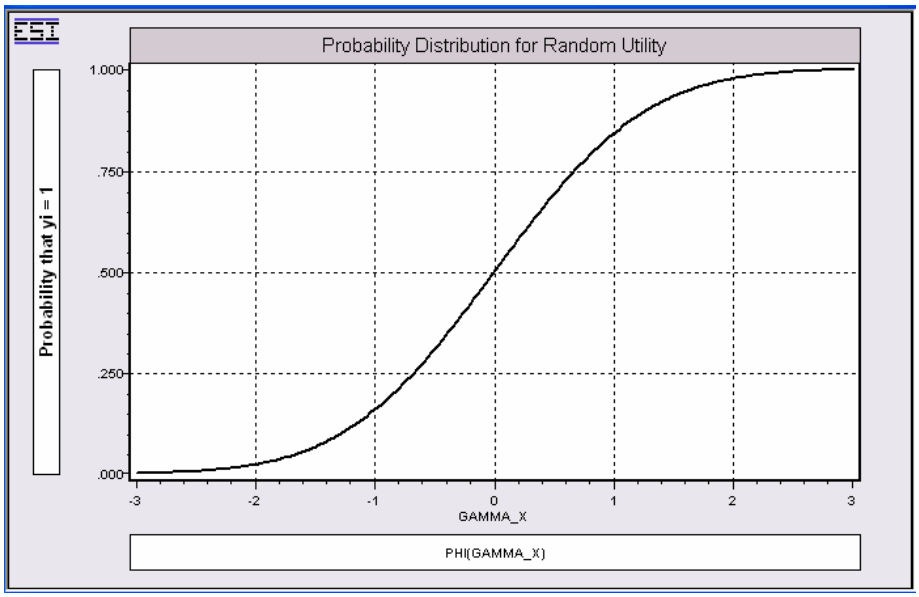


Figure 2.3 Probit Model for Binary Choice

There is an issue of identification in the binary choice model. We have assumed that the random term in the random utility function has a zero mean and known variance equal to one for

the normal distribution and $\pi^2/3$ for the logistic. These are *normalizations* of the model. Consider the zero mean assumption first. Assume that rather than having mean zero, ε_i has nonzero mean θ. The model for determination of y_i will then be

$$\begin{aligned}\text{Prob}(y_i = 1|\mathbf{x}_i) &= \text{Prob}(\varepsilon_i < \alpha + \tilde{\boldsymbol{\gamma}}'\mathbf{x}_i) \\ &= \text{Prob}(\varepsilon_i - \theta < (\alpha - \theta) + \tilde{\boldsymbol{\gamma}}'\mathbf{x}_i) \\ &= \text{Prob}(\varepsilon_i^* < \alpha^* + \tilde{\boldsymbol{\gamma}}'\mathbf{x}_i).\end{aligned}$$

Where $\tilde{\boldsymbol{\gamma}}$ is the rest of $\boldsymbol{\gamma}$ not including the constant term. The same model results, with ε_i^* now having a zero mean and the nonzero mean of ε_i being absorbed into the constant term of the original model. The end result is that as long as the binary choice model contains a constant term, there is no loss of generality in assuming the mean of the random term is zero. A nonzero mean would disappear into the constant term of the utility function. The reason for the assumption of a known variance is more subtle. Suppose that ε_i comes from population with standard deviation σ. For convenience, write $\varepsilon_i = \sigma v_i$ where v_i has zero mean and standard deviation one. Then,

$$y_i = 1[\boldsymbol{\gamma}'\mathbf{x}_i + \sigma v_i > 0].$$

Now, multiply the term in square brackets by any positive constant, λ. The same observation mechanism results; because we only observe zeros and ones,

$$\begin{aligned}y_i &= 1[\lambda(\boldsymbol{\gamma}'\mathbf{x}_i + \sigma v_i) > 0)] \\ &= 1[\boldsymbol{\gamma}^{*\prime}\mathbf{x}_i + \sigma^* v_i > 0],\end{aligned}$$

for any positive λ we might choose. We can assume any positive σ and observe exactly the same data, the same zeros and ones. Contrast this to the linear regression model,

$$y_i = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i,$$

in which a scaling of the right hand side of the equation translates into an equal scaling of y_i. To remove the indeterminacy in the probit model, it is conventional to assume that $\sigma = 1$. In the logit model, $f(\varepsilon_i)$ is kept in the standardized form with implied standard deviation, $\sigma = \pi/\sqrt{3}$. The end result is that because y_i has no scale – it is always zeros and ones – the data do not provide any way that we could estimate a variance parameter.

2.3 Estimation and Inference

Estimation and inference for probit and logit models for binary choice models is usually based on maximum likelihood estimation. The recent literature does contain some applications of Bayesian methods, so we will examine a Bayesian estimator as well.

2.3.1 Maximum Likelihood Estimation

Each observation is a draw from a Bernoulli distribution (binomial with one trial). The model with success probability $F(\boldsymbol{\gamma}'\mathbf{x}_i)$ and independent observations leads to the joint probability, or *likelihood function*,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{y_i=0}[1 - F(\boldsymbol{\gamma}'\mathbf{x}_i)] \times \prod_{y_i=1} F(\boldsymbol{\gamma}'\mathbf{x}_i). \quad (2.8)$$

Let **X** denote the sample of n observations, where the ith row of **X** is the ith observation on $\mathbf{x}_i$ (transposed, since $\mathbf{x}_i$ is a column) and let **y** denote the column vector that is the n observations on y_i. Then, the likelihood function for the parameters may be written

$$L(\boldsymbol{\gamma}|\mathbf{X},\mathbf{y}) = \prod_{i=1}^{n} [1-F(\boldsymbol{\gamma}'\mathbf{x}_i)]^{1-y_i}[F(\boldsymbol{\gamma}'\mathbf{x}_i)]^{y_i}. \tag{2.9}$$

Taking logs, we obtain the *log likelihood function*,

$$\ln L(\boldsymbol{\gamma}|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n} (1-y_i)\ln[1-F(\boldsymbol{\gamma}'\mathbf{x}_i)] + y_i \ln F(\boldsymbol{\gamma}'\mathbf{x}_i). \tag{2.10}$$

We are limiting our attention to the normal and logistic, symmetric distributions. This permits a useful simplification. Let

$$q_i = 2y_i - 1. \tag{2.11}$$

Thus, q_i equals -1 when y_i equals zero and $+1$ when y_i equals one. Because the symmetric distributions have the property that $F(t) = 1 - F(-t)$, we can combine the preceding into

$$\ln L(\boldsymbol{\gamma}|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n} \ln F[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]. \tag{2.12}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\gamma}$ is the vector of values that maximizes this function.

The MLE is the solution to the *likelihood equations*,

$$\frac{\partial \ln L(\boldsymbol{\gamma} \mid \mathbf{X},\mathbf{y})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left\{ q_i \frac{f[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}{F[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]} \right\} \mathbf{x}_i = \mathbf{0}, \tag{2.13}$$

where $f(.)$ is the density, $dF(.)/d(\boldsymbol{\gamma}'\mathbf{x}_i)$. The likelihood equations will be nonlinear and require an iterative solution. For the logit model, the likelihood equations can be reduced to

$$\frac{\partial \ln L(\boldsymbol{\gamma} \mid \mathbf{X},\mathbf{y})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left[y_i - \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i) \right] \mathbf{x}_i = \mathbf{0}. \tag{2.14}$$

If $\mathbf{x}_i$ contains a constant term, then, by multiplying the likelihood equation by $1/n$, we find that the first-order condition with respect to the constant term implies

$$\frac{1}{n}\sum_{i=1}^{n} \left[y_i - \Lambda(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i) \right] = 0. \tag{2.15}$$

where $\hat{\boldsymbol{\gamma}}$ is the MLE of $\boldsymbol{\gamma}$. That is, the average of the predicted probabilities must equal the proportion of ones in the sample, $P_1 = (1/n)\Sigma_i y_i$. Although the same result has not been shown to hold exactly (theoretically) for the probit model, it does appear as a striking empirical regularity there as well. The likelihood equation also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)$ as a residual. The first derivative of the log-likelihood

with respect to the constant term produces the *generalized residual* in many settings. [See, for example, Chesher and Irish (1985).] The log-likelihood function for the probit model is

$$\begin{aligned}\ln L(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{y}) &= \sum_{y_i=0}^{n} \ln[1-\Phi(\boldsymbol{\gamma}'\mathbf{x}_i)] + \sum_{y_i=1}^{n} \ln \Phi(\boldsymbol{\gamma}'\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \ln \Phi[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)].\end{aligned} \tag{2.16}$$

The likelihood equations are

$$\begin{aligned}\frac{\partial \ln L(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{y})}{\partial \boldsymbol{\gamma}} &= \sum_{y_i=0} \left[\frac{-\phi(\boldsymbol{\gamma}'\mathbf{x}_i)}{1-\Phi(\boldsymbol{\gamma}'\mathbf{x}_i)}\right]\mathbf{x}_i + \sum_{y_i=1} \left[\frac{\phi(\boldsymbol{\gamma}'\mathbf{x}_i)}{\Phi(\boldsymbol{\gamma}'\mathbf{x}_i)}\right]\mathbf{x}_i \\ &= \sum_{y_i=0} \lambda_i^0 \mathbf{x}_i + \sum_{y_i=1} \lambda_i^1 \mathbf{x}_i \\ &= \sum_{i=1}^{n} \left[q_i \frac{\phi[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}{\Phi[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}\right]\mathbf{x}_i \\ &= \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \\ &= \mathbf{0}.\end{aligned} \tag{2.17}$$

Note that λ_i is negative when y_i equals zero and positive when y_i equals one.

2.3.2 Maximizing the Log Likelihood Function

The second derivatives of the log likelihood function are

$$\begin{aligned}\mathbf{H} &= \frac{\partial^2 \ln L(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{y})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \\ &= \sum_{i=1}^{n} \left\{\frac{f'[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}{F[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]} - \left(\frac{f[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}{F[q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]}\right)^2\right\} q_i^2 \mathbf{x}_i \mathbf{x}_i'.\end{aligned} \tag{2.18}$$

Where $f'(t)$ is the derivative of the density function for the normal or logistic distribution. For the normal distribution, this is $\phi'(t) = -t\phi(t)$ while for the logistic distribution, this is

$$f'(t) = [1-2\Lambda(t)]\Lambda(t)[1-\Lambda(t)]. \tag{2.19}$$

These expressions simplify the second derivatives considerably. For the probit model,

$$\mathbf{H}_P = \sum_{i=1}^{n} \left\{-\lambda_i[\lambda_i + q_i(\boldsymbol{\gamma}'\mathbf{x}_i)]\right\} \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} h_{i,P} \mathbf{x}_i \mathbf{x}_i' \tag{2.20a}$$

while for the logit model this is

$$\mathbf{H}_L = \sum_{i=1}^{n} \left\{-\Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)[1-\Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)]\right\} \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} h_{i,L} \mathbf{x}_i \mathbf{x}_i'. \tag{2.20b}$$

In both of these cases, the term in braces, $h_{i,Model}$, is always negative. This means that the second derivatives matrix of the log likelihood is always negative definite, which greatly simplifies maximization of the function.

Newton's method uses the iteration

$$\hat{\boldsymbol{\gamma}}(r+1) = \hat{\boldsymbol{\gamma}}(r) - \left[\hat{\mathbf{H}}(r)\right]^{-1} \hat{\mathbf{g}}(r), \tag{2.21}$$

where r indexes the iterations, $\hat{\mathbf{H}}(r)$ is the second derivatives matrix computed at the current value of the parameters and $\hat{\mathbf{g}}(r)$ is the vector of first derivatives evaluated at the current values of the parameters. An alternative method based on the expected value of the second derivatives matrix is the *method of scoring*. The Hessian for the logit model is not a function of y_i (i.e., q_i), so

$E[\mathbf{H}_L] = \mathbf{H}_L$.

For the probit model, a considerable amount of tedious algebra produces the result

$$E[\mathbf{H}_P] = \sum_{i=1}^{n} \left\{ \frac{-\left[\phi(\boldsymbol{\gamma}'\mathbf{x}_i)\right]^2}{\Phi(\boldsymbol{\gamma}'\mathbf{x}_i)\left[1-\Phi(\boldsymbol{\gamma}'\mathbf{x}_i)\right]} \right\} \mathbf{x}_i\mathbf{x}_i'. \tag{2.22}$$

The method of scoring is used by replacing $\mathbf{H}(r)$ with $E[\mathbf{H}(r)]$ in Newton's method. Because of the shape of the log likelihood function – the negative definiteness of the Hessian implies that the function is globally concave; it has only one mode – maximization using either of these methods is likely to be fast and simple. [See Pratt (1981).]

Two other methods of maximizing the log likelihod are interesting to examine at this point, the *EM algorithm* and a Bayesian estimator, the *Markov Chain Monte Carlo* (MCMC) approach using a *Gibbs sampler*. Neither of these methods is well suited to estimation of the logit model, surprisingly for the same reason. In each case, the mean of the truncated random variable, $E[\varepsilon_i | \varepsilon_i > -\boldsymbol{\gamma}'\mathbf{x}_i]$ is needed. The result is well known for the probit model but not for the logit model.

2.3.3 The EM Algorithm

The EM method is built around the idea that the probit model is a *missing data model*. If $U_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i$ were observed, the estimation problem would be much simpler; $\boldsymbol{\gamma}$ would be estimated by a linear regression of U_i^* on $\mathbf{x}_i$. With the normality assumption, this would be the maximum likelihood estimator. To use the EM algorithm, we would maximize the log likelihood function that is constructed by replacing U_i^* with $\mathrm{E}[U_i^* | y_i, \mathbf{x}_i]$. [The method is only equivalent to doing this regression – see Dempster, Laird and Rubin (1977) for the actual specifics of the algorithm. We will also add some details in Section 8.2.3.] The conditional mean functions we need are

$$\begin{aligned} E[U_i^* | y_i = 1, \mathbf{x}_i] &= E[\boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i \mid \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i > 0, \mathbf{x}_i] \\ &= \boldsymbol{\gamma}'\mathbf{x}_i + \mathrm{E}[\varepsilon_i \mid \varepsilon_i > -\boldsymbol{\gamma}'\mathbf{x}_i] \\ &= \boldsymbol{\gamma}'\mathbf{x}_i + \frac{\phi(-\boldsymbol{\gamma}'\mathbf{x}_i)}{1-\Phi(-\boldsymbol{\gamma}'\mathbf{x}_i)} \\ &= \boldsymbol{\gamma}'\mathbf{x}_i + \lambda_i^1. \end{aligned} \tag{2.23a}$$

[See (2.17).] By the same logic, then

$$E[U_i^* \mid y_i = 0, \mathbf{x}_i] \;=\; \boldsymbol{\gamma}'\mathbf{x}_i + \lambda_i^0 . \tag{2.23b}$$

The iteration works as follows: We begin with a starting value of $\boldsymbol{\gamma}$, say $\boldsymbol{\gamma}(0)$. At each iteration, r, we compute the predictions

$$\hat{U}_i^*(r) = \hat{\boldsymbol{\gamma}}(r)'\mathbf{x}_i + \hat{\lambda}_i(r).$$

Then, the new $\hat{\boldsymbol{\gamma}}(r+1)$ is computed by linear regression of $\hat{U}_i^*(r)$ on $\mathbf{x}_i$. Therefore, the iteration for the EM method is

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(r+1) &= \left[\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \sum_{i=1}^n \mathbf{x}_i \hat{U}_i^*(r) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^n \mathbf{x}_i\left[\hat{\boldsymbol{\gamma}}(r)'\mathbf{x}_i + \hat{\lambda}_i(r)\right] \\ &= \hat{\boldsymbol{\gamma}}(r) + (\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^n \mathbf{x}_i\hat{\lambda}_i(r).\end{aligned} \tag{2.24}$$

Notice that the summation at the end is just the derivatives of the log likelihood function evaluated at $\hat{\boldsymbol{\gamma}}(r)$. [See (2.17).] This means that the EM method for the probit model is the same as Newton's method or the method of scoring except that $(\mathbf{X}'\mathbf{X})^{-1}$ is used in place of $-\mathbf{H}$ or $-E[\mathbf{H}]$. For this particular model (not in general), the EM method is not a particularly effective approach to maximizing the log likelihood function. Using $(\mathbf{X}'\mathbf{X})^{-1}$ instead of the Hessian in a Newton-like iteration turns out generally to be a slower method of maximizing the log likelihood. There are fewer computations, since $(\mathbf{X}'\mathbf{X})^{-1}$ needs ony to be computed ones. But, typically, many more iterations than Newton's method are required to locate the solution.

2.3.4 Bayesian Estimation by Gibbs Sampling and MCMC

Bayesian estimation of a probit model builds on the method pioneered by Albert and Chib (1993). [See Lancaster (2004) for this development.] The Gibbs sampler is constructed using a crucial device labeled *data augmentation*. [See Tanner and Wong (1987).] The binary choice case departs from

$$y_i^* \;=\; \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i,\;\; \varepsilon_i \sim \text{ with mean 0 and known variance, 1 (probit) or } \pi^2/3 \text{ (logit)},$$

$$y_i \;=\; 1 \;\text{ if } y_i^* > 0.$$

Let the prior for $\boldsymbol{\gamma}$ be denoted $p(\boldsymbol{\gamma})$. Then, the *posterior density* for the probit or logit (symmetric distribution) models is

$$p(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) = \frac{p(\boldsymbol{\gamma})\prod_{i=1}^n F[q_i\boldsymbol{\gamma}'\mathbf{x}_i]}{\int_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma})\prod_{i=1}^n F[q_i\boldsymbol{\gamma}'\mathbf{x}_i]d\boldsymbol{\gamma}}, \tag{2.25}$$

where we use $\mathbf{y}$ and $\mathbf{X}$ (and later, $\mathbf{y}^*$) to denote the full set of N observations. Estimation of the posterior mean is done by setting up a Gibbs sampler in which the unknown values y_i^* are treated

as nuisance parameters to be estimated along with $\boldsymbol{\gamma}$. For convenience at this point, we will assume the probit model is of interest. Conditioned on $\boldsymbol{\gamma}$ and $\mathbf{x}_i$, y_i^* has a normal distribution with mean $\boldsymbol{\gamma}'\mathbf{x}_i$ and variance 1. However, when conditioned on y_i (observed), as well, the sign of y_i^* is known;

$$p(y_i^* \mid \boldsymbol{\gamma},\mathbf{y},\mathbf{X}) = \text{normal with mean } \boldsymbol{\gamma}'\mathbf{x}_i \text{ and variance 1, truncated at zero;}$$
$$\text{truncated from below if } y_i = 1 \text{ and from above if } y_i = 0.$$

Using basic results for Bayesian analysis of the linear model with known disturbance [see Greene (2008a, p. 605)] and a diffuse prior, the posterior for $\boldsymbol{\gamma}$ conditioned on $\mathbf{y}^*$, $\mathbf{y}$ and $\mathbf{X}$ would be

$$p(\boldsymbol{\gamma}\mid \mathbf{y}^*,\mathbf{y},\mathbf{X}) = N_K[\mathbf{c},(\mathbf{X}'\mathbf{X})^{-1}] \text{ where } \mathbf{c} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*.$$

If, instead, the prior for $\boldsymbol{\gamma}$ is normal with mean $\boldsymbol{\gamma}^0$ and covariance matrix, $\boldsymbol{\Sigma}$, then the posterior density is normal with *posterior mean*

$$E[\boldsymbol{\gamma}|\mathbf{y}^*,\mathbf{y},\mathbf{X}] = [\boldsymbol{\Sigma}^{-1} + (\mathbf{X}'\mathbf{X})]^{-1}\,(\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\gamma}^0 + \mathbf{X}'\mathbf{y}^*)$$

and

$$\text{Var}[\boldsymbol{\gamma}|\mathbf{y}^*,\mathbf{y},\mathbf{X}] = [\boldsymbol{\Sigma}^{-1} + (\mathbf{X}'\mathbf{X})]^{-1}.$$

This sets up a simple Gibbs sampler for drawing from the joint posterior, $p(\boldsymbol{\gamma},\mathbf{y}^*|\mathbf{y},\mathbf{X})$. It is customary to use a diffuse prior for $\boldsymbol{\gamma}$. Then, compute initially, $(\mathbf{X}'\mathbf{X})^{-1}$ and the lower triangular *Cholesky matrix*, $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}' = (\mathbf{X}'\mathbf{X})^{-1}$. (The matrix $\mathbf{L}$ needs only to be computed once at the outset for the informative prior as well. In that case, $\mathbf{L}\mathbf{L}' = (\boldsymbol{\Sigma}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$.) To initialize the iterations, any reasonable value of $\boldsymbol{\gamma}$ may be used. Albert and Chib suggest the classical MLE. The iterations are then given by

1. Compute the N draws from $p(\mathbf{y}^*|\boldsymbol{\gamma},\mathbf{y},\mathbf{X})$.
   Draws from the appropriate truncated normal can be obtained using

   $$y_i^*(r) = \boldsymbol{\gamma}'\mathbf{x}_i + \Phi^{-1}[\Phi(-\boldsymbol{\gamma}'\mathbf{x}_i) + U \times (1-\Phi(-\boldsymbol{\gamma}'\mathbf{x}_i))] \text{ if } y_i = 1 \text{ and}$$

   $$y_i^*(r) = \boldsymbol{\gamma}'\mathbf{x}_i + \Phi^{-1}[U \times \Phi(-\boldsymbol{\gamma}'\mathbf{x}_i)] \text{ if } y_i = 0,$$

   where U is a single draw from a standard uniform population and $\Phi^{-1}(U)$ is the inverse function of the standard normal.
2. Draw an observation on $\boldsymbol{\gamma}$ from the posterior $p(\boldsymbol{\gamma}|\mathbf{y}^*,\mathbf{y},\mathbf{X})$ by first computing the mean

   $$\mathbf{c}(r) = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y}^*(r).$$

   Use a draw, $\mathbf{v}$, from the K-variate standard normal, then compute $\boldsymbol{\gamma}(r) = \mathbf{c}(r) + \mathbf{L}\mathbf{v}$.

(We have used "(r)" to denote the rth cycle of the iteration.) The iteration cycles between steps 1 and 2 until a satisfactory number of draws is obtained (and a burn-in number are discarded), then the retained observations on $\boldsymbol{\gamma}$ are analyzed. With an informative prior, the draws at step 2 involving the prior mean and variance are slightly more time consuming. The matrix $\mathbf{L}$ is only computed at the outset, but the computation of the mean adds a matrix multiplication and addition.

The MCMC estimator for this model shows an interesting application of the technique. For the probit model, in particular, however, it can be an extremely inefficient method of

estimation. With a diffuse prior, the posterior density will look very much like log likelihood and, particularly when the sample is reasonably large, the posterior mean will be essentially the same as the MLE. But, estimating the posterior mean will require possibly thousands of generations of thousands of observations (on $y_i^*(r)$) each followed by a regression, compared to a small handful of regressions for Newton's method. In a sample of two thousand, for example, we found that the MCMC estimator took more than 25 times as long as Newton's method to reach essentially the same set of results.

We have developed the Bayesian estimator in this section to illustrate the technique and to introduce a few concepts, including the Gibbs sampler and the method of data augmentation which is extremely useful in discrete choice modeling. Save for a few applications to be presented later, we will now focus on classical, likelihood based methods of estimation and inference.

2.3.5 Estimation with Grouped Data and Iteratively Reweighted Least Squares

Many applications of binary choice modeling in biological and social sciences involve *grouped data*. Consider, for example, a study intended to learn the appropriate dosage level of a drug or the effectiveness of a pesticide. A group of n_i individuals is subjected to dosage level x_i and a proportion $p_{i1} = n_{i1}/n_i$ respond to the drug (by recovering or by dying). Thus, proportion $p_{i0} = 1 - p_{i1}$ do not respond. The term *repeated measures* is sometimes applied to such data. This setting is only slightly different from the one we have examined so far. Let Y_i equal the number of responders among the n_i subjects and let y_{it} denote whether individual t in group i responds. Then $Y_i = \Sigma_t y_{it}$. We assume that the random utility/binary choice model applies to each subject, where y_{it}^* would correspond to the subject's own tolerance or resistance level to the treatment.

The log likelihood function would be

$$\begin{aligned}
\ln L &= \sum_{i=1}^{N} \left[\sum_{y_{it}=0} \ln F(-\boldsymbol{\gamma}'\mathbf{x}_i) + \sum_{y_{it}=1} \ln F(\boldsymbol{\gamma}'\mathbf{x}_i) \right] \\
&= \sum_{i=1}^{N} n_{i0} \ln F(-\boldsymbol{\gamma}'\mathbf{x}_i) + n_{i1} \ln F(\boldsymbol{\gamma}'\mathbf{x}_i) \\
&= \sum_{i=1}^{N} n_i \left[p_{i0} \ln F(-\boldsymbol{\gamma}'\mathbf{x}_i) + p_{i1} \ln F(\boldsymbol{\gamma}'\mathbf{x}_i) \right] \\
&= \sum_{i=1}^{N} n_i \left[(1-p_i) \ln(1-F_i) + p_i \ln F_i \right].
\end{aligned} \tag{2.26}$$

This is the same function we maximized earlier, where $n_i = 1$, $p_{i0} = 1\text{-}y_i$ and $p_{i1} = y_i$.

Johnson and Albert (1999) note that this function can be maximized by a type of *iteratively reweighted least squares*. The authors argue that "*The difficulty in obtaining maximum likelihood estimates for a binary regression stems from the complexity of (3.16), which makes an analytic expression for the maximum likelihood estimates of* $\boldsymbol{\beta}$ *impossible to obtain. However, the iteratively reweighted least squares algorithm makes point estimation of maximum likelihood estimates a trivial task and underlies the algorithm used in most commercial software packages.*" (p. 119). The iteratively weighted least squares method was pioneered by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) for the class of *generalized linear models*. For the case of a binary regression model, the technique is simply another Newton-like method, as we now demonstrate.

Using our notation but the identical functions, Johnson and Albert (1999) define the algorithm as follows: Let

$$z_i = \boldsymbol{\gamma}'\mathbf{x}_i + \frac{(Y_i - n_i F_i)}{n_i dF_i / d(\boldsymbol{\gamma}'\mathbf{x}_i)}$$
$$= \boldsymbol{\gamma}'\mathbf{x}_i + \frac{(p_i - F_i)}{f_i},$$

where $\boldsymbol{\gamma}$ is the current estimate of the parameter vector, Y_i is the number of responders in group i, F_i is the probability (logit or probit cdf), f_i is the density (derivative of F_i) and the second line is obtained by noting that $Y_i = n_i p_i$. Define the weight,

$$w_i = \frac{n_i f_i^2}{F_i(1-F_i)}.$$

The iteratively reweighted least squares estimator is obtained by weighted least squares regression of z_i on $\mathbf{x}_i$, with weights w_i. Thus, the iterative estimator is

$$\boldsymbol{\gamma}^1 = \left[\sum_{i=1}^n w_i^0 \mathbf{x}_i \mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^n w_i^0 \mathbf{x}_i z_i^0\right],$$

where the terms w_i^0 and z_i^0 are computed using $\boldsymbol{\gamma}^0$. It is obvious based on the form of z_i^0 that this can be written

$$\boldsymbol{\gamma}^1 = \boldsymbol{\gamma}^0 + \left[\sum_{i=1}^n w_i^0 \mathbf{x}_i \mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^n \mathbf{x}_i w_i^0 \frac{(p_i - F_i)}{f_i}\right].$$

The derivative of the log likelihood with respect to $\boldsymbol{\gamma}$ is, after a bit of manipulation,

$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^N \mathbf{x}_i \frac{n f_i (p_i - F_i)}{F_i(1-F_i)}.$$

By multiplying w_i^0 by $(p_i - F_i)/f_i$ in the iteration, we find the product exactly equals the scalar term in the derivative. It follows, then, that the iteratively reweighted least squares estimator is simply

$$\boldsymbol{\gamma}^1 = \boldsymbol{\gamma}^0 + \left[\sum_{i=1}^n w_i^0 \mathbf{x}_i \mathbf{x}_i'\right]^{-1}\left(\frac{\partial \ln L^0}{\partial \boldsymbol{\gamma}^0}\right).$$

Ostensibly, the only difference between this and Newton's method is the weighting matrix in brackets. However, for the logit model, $f_i = F_i(1\text{-}F_i)$, from which it follows that the estimator is, in fact, identical to Newton's method. For the probit model it differs slightly because of the form of f_i; but it is surely no more complicated to compute.

2.3.6 The Minimum Chi Squared Estimator

The *minimum chi squared estimator* (MCS) is obtained by treating p_{i1} as an estimator, subject to sampling variability, of $F(\boldsymbol{\gamma}'\mathbf{x}_i)$. The simpler case to work with is the logit model. Write

$p_{i1} = F(\boldsymbol{\gamma}'\mathbf{x}_i) + w_i.$

Then, $\log[p_{i1}/(1-p_{i1})] = \text{logit}(p_{i1}) = \boldsymbol{\gamma}'\mathbf{x}_i + w_i^*$ where w_i^* is a heteroscedastic error of approximation that embodies w_i and the error in linearization of the function. For the logit model, $\boldsymbol{\gamma}$ may now be estimated by weighted least squares regression of $\text{logit}(p_{i1})$ on $\mathbf{x}_i$ with weights $1/\left[n_i p_{i1}(1-p_{i1})\right]$. The estimator may be iterated by replacing p_{i1} in the weights with $\hat{F}_i$ from the previous iteration. It has been shown that the MCS estimator, though numerically different, is as efficient as MLE. [See Greene (2003, pp. 688-689).] Nonetheless, the MLE is the preferred estimator in nearly all contemporary applications.

2.4 Covariance Matrix Estimation

There are three available estimators for the asymptotic covariance matrix of the MLE. The conventional approach is based on the actual second derivatives of the log likelihood;

$$\begin{aligned} Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}_{MLE}\right] &= \left[-\mathbf{H}\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\right]^{-1} \\ &= \left[-\sum_{i=1}^{n} h_i\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}, \end{aligned}$$

where the expressions for h_i are given in the braces in the expressions for $\mathbf{H}_P$ and $\mathbf{H}_L$ in (2.20). A second approach that is usually not available for more complicated models, but is for the probit and logit models, is to base the covariance matrix estimation on the expected Hessian, rather than the actual estimated one. This is the same matrix for the logit model. For the probit model, the estimator is

$$\begin{aligned} Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}_{P,MLE}\right] &= \left[\sum_{i=1}^{n} \left\{\frac{\left[\phi(\hat{\boldsymbol{\gamma}}_{P,MLE}'\mathbf{x}_i)\right]^2}{\Phi(\hat{\boldsymbol{\gamma}}_{P,MLE}'\mathbf{x}_i)\left[1-\Phi(\hat{\boldsymbol{\gamma}}_{P,MLE}'\mathbf{x}_i)\right]}\right\}\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \\ &= \left[\sum_{i=1}^{n} -\hat{\lambda}_i^0\hat{\lambda}_i^1\mathbf{x}_i\mathbf{x}'\right]^{-1}. \end{aligned} \tag{2.27}$$

The third estimator is the Berndt, Hall, Hall and Hausman (1973) estimator based only on the first derivatives;

$$Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}_{MLE}\right]_{BHHH} = \left[\sum_{i=1}^{n} g_i^2\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}, \tag{2.28}$$

where the first derivative terms are $g_i = [y_i - \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)]$ for the logit model and $g_i = \lambda_i$ for the probit model. Any of the three of these may be used; all are appropriate estimators of the asymptotic covariance matrix for the MLE.

Robust Covariance Matrix Estimation

The probit and logit maximum likelihood estimators are often labeled *quasi-maximum likelihood estimators* (QMLE) in view of the possibility that the normal or logistic probability model might be misspecified. White's (1982a) robust "sandwich" estimator for the asymptotic covariance matrix of the QMLE,

$$
\begin{aligned}
&Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}_{MLE}\right] \\
&= \left[-\sum_{i=1}^{n} h_i\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n} g_i^2\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\mathbf{x}_i\mathbf{x}_i'\right]\left[-\sum_{i=1}^{n} h_i\left(\hat{\boldsymbol{\gamma}}_{MLE}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \qquad (2.29)\\
&= (-\hat{\mathbf{H}})^{-1}\hat{\mathbf{B}}(-\hat{\mathbf{H}})^{-1},
\end{aligned}
$$

has been used in a number of studies based on the probit model [e.g., Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993)] and is a standard feature in contemporary software such as *Stata* and *NLOGIT*.

If the probit model is correctly specified, then plim$(1/n)\,\hat{\mathbf{B}}$ = plim$(1/n)(-\hat{\mathbf{H}})$ and either single matrix will suffice, so the robustness issue is moot. But, the probit (Q-) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), correlation across observations, nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction. [See Greene (2008a, Section 16.8) and Freedman (2006).] White raises this issue explicitly, although it seems to receive little attention in the literature: "*It is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the basis for robust estimation techniques*" (1982a, p. 4). His very useful result is that if the quasi-maximum likelihood estimator converges to a probability limit, then the sandwich estimator can, under certain circumstances, be used to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear.

2.5 Application of the Binary Choice Model to Health Satisfaction

Riphahn, Wambach and Million (RWM, 2003) analyzed individual data on health care utilization (doctor visits and hospital visits) using various models for counts. The data set is a large panel extracted from the German Socioeconomic Panel (GSOEP). [See RWM (2003) and Greene (2008a) for discussion of the data set in detail.] The data set is an unbalanced panel including 7,293 German households observed from 1 to 7 times and a total of 27,326 observations. (We will visit the panel data aspects of the data and models later.) Among the several interesting variables in this data set is *HSAT*, a self reported health assessment that is recorded with values 0,1,..,10. The sample mean response is 6.8. To construct an example for this chapter, we will define the dependent variable

$$
Healthy_i = \begin{array}{l} 1 \text{ if } HSAT_i \geq 7 \\ 0 \text{ otherwise.} \end{array}
$$

The families were observed in 1984-1988, 1991 and 1995. For purposes of the application, to maintain as closely as possible the assumptions of the model, at this point, we have selected the most frequently observed year, 1988, for which there are a total of 4,483 observations, 2,313 males (*Female* = 0) and 2,170 females (*Female* = 1). We will use the following variables in the regression part of the model,

$$
\mathbf{x}_i = (constant,\ Age_i,\ Income_i,\ Education_i,\ Married_i,\ Kids_i).
$$

In the original data set, *Income* is *HHNINC* (household income) and *Kids* is *HHKIDS* (household kids). *Married* and *Kids* are binary variables, the latter indicating whether or not there are children in the household. Descriptive statistics for the data used in the application are shown in Table 2.1.

Table 2.1 Data Used in Binary Choice Application

Variable	Male (n=2313) Mean	S.D.	Female (n=2170) Mean	S.D.	All (n=4483) Mean	S.D.	Min.	Max.
HEALTHY	0.624	0.485	0.582	0.493	0.604	0.489	0	1
AGE	42.73	11.39	44.20	11.23	43.44	11.29	25	64
EDUC	11.83	2.494	10.98	2.142	11.42	2.368	7	18
INCOME	0.355	0.165	0.342	0.163	0.349	0.164	0	2
MARRIED	0.756	0.429	0.748	0.434	0.752	0.432	0	1
KIDS	0.387	0.487	0.372	0.483	0.379	0.485	0	1

Estimates of the parameters of the probit and logit models are shown in Table 2.2. In terms of the diagnostic statistics, the log likelihood function and the *t* ratios for the parameters, the two models appear almost identical. However, there are prominent differences between the coefficients. To a reasonable approximation, the regularity, that will show up in most cases, is

$$\frac{\hat{\gamma}_{k,LOGIT}}{\hat{\gamma}_{k,PROBIT}} \approx 1.6 .$$

The substantial difference between the coefficients in the two models exaggerates the substantive difference between the specifications. When we turn, instead, to the partial effects implied by the models, the difference largely disappears. An example appears below. Table 2.3 displays the standard errors obtained by the different methods shown earlier. As might be expected in a sample this large, and in the absence of some major flaw in the model specification, the estimates are almost identical. Note that this holds even for the "robust" estimator. We have only shown the results for the probit model, but they are almost identical for the logit model.

Table 2.2 Estimated Probit and Logit Models

	Logit LogL = -2890.393				Probit LogL = -2890.288				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.7595	.2349	3.233	.0012	.4816	.1423	3.383	.0007	1.0000
AGE	-.0329	.0032	-10.266	.0000	-.0203	.0020	-10.386	.0000	43.4401
EDUC	.0860	.0148	5.805	.0000	.0520	.0089	5.872	.0000	11.4181
INCOME	.3454	.2083	1.658	.0972	.2180	.1265	1.724	.0847	.34874
MARRIED	-.0483	.0828	-.584	.5592	-.0311	.0508	-.612	.5403	.75217
KIDS	.1278	.0756	1.692	.0907	.0800	.0463	1.727	.0841	.37943

Table 2.3 Alternative Estimated Standard Errors for the Probit Model

Variable	Coefficient	Std.Error E[H]	Std.Error H	Std. Error BHHH	Std.Error Robust
Constant	.48160673	.14234358	.14248075	.14191210	.14282907
AGE	-.02035358	.00195967	.00196013	.00195847	.00196118
EDUC	.05204356	.00886339	.00890657	.00870141	.00902935
INCOME	.21801803	.12646534	.12695827	.12469930	.12830838
MARRIED	-.03107496	.05075075	.05081510	.05053493	.05098501
KIDS	.08004423	.04633938	.04629982	.04649873	.04618606

2.6 Partial Effects in a Binary Choice Model

The probability model is a regression:

$$E[y_i \mid \mathbf{x}_i] = 0 \times [1 - F(\boldsymbol{\gamma}'\mathbf{x}_i)] + 1 \times F(\boldsymbol{\gamma}'\mathbf{x}_i) = F(\boldsymbol{\gamma}'\mathbf{x}_i).$$

Therefore, the probability that y_i equals one is also the expectation, or regression function. Whatever distribution is used, it is important to note that the parameters of the model, like those of any nonlinear regression model, are not necessarily the *marginal effects* we are accustomed to analyzing. In general,

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \left[\frac{dF(\boldsymbol{\gamma}'\mathbf{x}_i)}{d(\boldsymbol{\gamma}'\mathbf{x}_i)}\right]\boldsymbol{\gamma}. \tag{2.30}$$

where $f(.)$ is the probability density function (PDF) that corresponds to the CDF, $F(.)$. For the normal distribution, this result is

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \phi(\boldsymbol{\gamma}'\mathbf{x}_i)\boldsymbol{\gamma}\,, \tag{2.31a}$$

where $\phi(t)$ is the standard normal density. For the logistic distribution,

$$\left[\frac{d\Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)}{d(\boldsymbol{\gamma}'\mathbf{x}_i)}\right] = \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)[1-\Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)]. \tag{2.32}$$

Thus, in the logit model,

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)[1-\Lambda(\boldsymbol{\gamma}'\mathbf{x}_i)]\boldsymbol{\gamma}. \tag{2.31b}$$

These values will vary with the values of $\mathbf{x}$. The effect is illustrated in Figure 2.4, where $\Delta x_k = 1$ for both cases, but $\Delta F(\boldsymbol{\gamma}'\mathbf{x})$ depends on where the calculation begins.. In interpreting the estimated model, it will be useful to calculate this value at, say, the means of the variables and, possibly, other specific values. For convenience, it is worth noting that the same scale factor applies to all the slopes in the model. For computing marginal effects**,** one can evaluate the expressions at the sample means of the data or evaluate the marginal effects at every observation and use the sample average of the individual marginal effects. Current practice favors averaging the individual marginal effects when it is possible to do so (though in practice, there is usually only minor numerical difference between the two results).

Recall, as shown in the earlier example, that the estimated coefficients in the logit model will generally be approximately equal to 1.6 times their counterparts in the probit model. The scale factors in (2.28a,b), at least near $\boldsymbol{\gamma}'\mathbf{x}_i = 0$, will be roughly 0.25 for the logit model and 0.4 for the probit model. Thus, the scaling to obtain the marginal effects will undo the difference in the coefficients. This effect is clearly visible in the example in Table 2.4.

2.6.1 Partial Effect for a Dummy Variable

Another complication for computing marginal effects in a binary choice model arises because **x** will often include dummy variables—for example, our application to health satisfaction includes dummy variables for marital status and number of children. Because the derivative is with respect to a small change, it is not appropriate to compute derivatives for the effect of a change in a dummy variable, or change of state. The appropriate marginal effect for a binary independent variable, say, d, would be

$$\text{Marginal effect} = [\text{Prob}(y_i = 1 | \bar{\mathbf{x}}_{(d)}, d_i = 1)] - [\text{Prob}(y_i = 1 | \bar{\mathbf{x}}_{(d)}, d_i = 0)],$$

where $\bar{\mathbf{x}}_{(d)}$, denotes the means of all the other variables in the model. Simply taking the derivative with respect to the binary variable as if it were continuous provides an approximation that is often surprisingly accurate. For example, the marginal effect for *Married* of -0.01191157 shown in table 2.4 is obtained by computing the probability that *Healthy* equals one while holding all the other variables at their means and *Married* equaling one then zero and taking the difference. The scale factor used to compute the partial effects for the other variables, using the values for *Income* in Tables 2.2 and 2.4 for the computation, is 0.08375583/0.21801803 = 0.3841693. Multiplying the coefficient on *Married* of -0.03107496 by this scale factor produces an estimated partial effect of -0.01193805. The error is only 0.2%. We will revisit this computation in the examples and discussions to follow. The first difference computation is now common in applications, and is built into modern software.

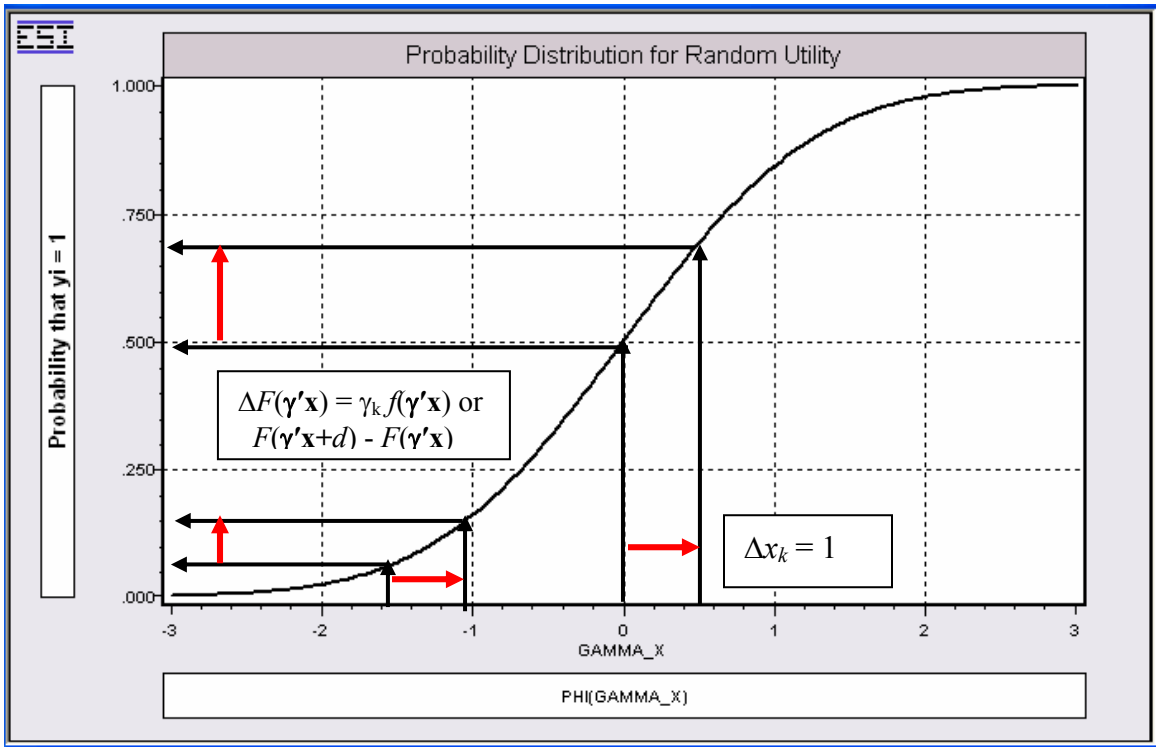


Figure 2.4 Partial Effects in a Binary Choice Model

The partial effects for the estimated probit and logit models are shown in Table 2.4. As expected, the different scaling of the two models that makes the coefficient estimates appear different is absent from the partial effects, which are nearly the same.

Table 2.4 Partial Effects for Probit and Logit Models at Means of x

	PROBIT				LOGIT			
Variable	Partial Effect	Std. Error	t	Elast.	Partial Effect	Std. Error	t	Elast.
AGE	-.0078	.0007	-10.39	-.5584	-.0078	.0006	-10.28	-.5587
EDUC	.0200	.0034	5.88	.3753	.0205	.0035	5.81	.3837
INCOME	.0838	.0486	1.72	.0480	.0822	.0496	1.66	.0471
MARRIED*	-.0119	.0194	-.61	-.0147	-.0115	.0196	-.59	-.0142
KIDS*	.0307	.0177	1.73	.0191	.0303	.0178	1.70	.0189

* Partial effects for dummy variables computed using discrete differences

2.6.2 Odds Ratios

For the logit model with a set of variables x and an additional (any) variable of interest, the odds in favor of a response of one are

$$\frac{\text{Prob}(y_i = 1 \mid \mathbf{x}, z)}{\text{Prob}(y_i = 0 \mid \mathbf{x}, z)} = \exp(\boldsymbol{\gamma}'\mathbf{x} + \theta z).$$

Consider how the the odds ratio changes when z changes by one unit

$$\frac{\left(\frac{\text{Prob}(y_i = 1 \mid \mathbf{x}, z+1)}{\text{Prob}(y_i = 0 \mid \mathbf{x}, z+1)}\right)}{\left(\frac{\text{Prob}(y_i = 1 \mid \mathbf{x}, z)}{\text{Prob}(y_i = 0 \mid \mathbf{x}, z)}\right)} = \exp(\theta) = \tau.$$

Analysts are occasionally interested in changes in odds ratios rather than changes in probabilities. Then, the interesting quantity to report as opposed to (or inaddition to) the partial effect is the change in the odds ratio, $\exp(\gamma_k)$. (*Stata* labels this as the "odds ratio" rather than the change in the odds ratio in its reported results.) Note that a full unit change in a variable is often not the change of interest. In our example, for instance, the income variable is scaled so that its full range of variation is only from zero to two, so a full unit change is not likely to be a useful measure for a derivative, even for an odds ratio with its ambiguous units of measurement. But, age, years of education and marital status are variables for which a one unit change is an empirically reasonable experiment.

2.6.3 Elasticities

It is common in some areas, such as transportation, to report elasticities of probabilities, rather than derivatives. These are straightforward to compute as

$$\begin{aligned}\varepsilon_{i,k} &= \frac{\partial \ln \text{Prob}(y_i = 1 \mid \mathbf{x}_i)}{\partial \ln x_{i,k}} \\ &= \frac{\partial \text{Prob}(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{i,k}} \frac{x_{i,k}}{\text{Prob}(y_i = 1 \mid \mathbf{x}_i)}.\end{aligned}$$

The elasticities are simple to obtain from the estimated partial effects. These are shown in Table 2.4 with the derivatives. We note, however, since it is a ratio of percentage changes, the elasticity is not likely to be useful for dummy variables such as marital status, or for discrete variables such as age and education.

Like a partial effect, an elasticity for a dummy variable or an integer valued variable will not necessarily produce a reasonable result. The computation for a dummy variable or an integer variable would be a semi-elasticity, [%ΔProb]/Δx, where Δx would equal one. Whether a percentage change in an integer valued x would make sense would depend on the context. Obviously it would not for a dummy variable. Whether it would for, say, years of education, would depend on the study in hand – we would surmise in most instances not, however. In sum, the relevant semi-elasticity for the change in a dummy variable (or a unit change in a discrete regressor) would be

$$e_{i,k} = \frac{\text{Prob}(y_i = j \mid \mathbf{x}_i, d_i = 1) - \text{Prob}(y_i = j \mid \mathbf{x}_i, d_i = 0)}{\frac{1}{2}\left[\text{Prob}(y_i = j \mid \mathbf{x}_i, d_i = 1) + \text{Prob}(y_i = j \mid \mathbf{x}_i, d_i = 0)\right]}$$

The denominator computation removes the asymmetry in the computation that makes it otherwise dependent on whether the change is from $d_i = 1$ or 0.

2.6.4 Inference for Partial Effects

The predicted probabilities, $F(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i) = \hat{F}_i$ and the estimated partial effects

$$\hat{\boldsymbol{\delta}} = f\left(\hat{\boldsymbol{\gamma}}'\bar{\mathbf{x}}\right)\hat{\boldsymbol{\gamma}},$$

are nonlinear functions of the parameter estimates. To compute standard errors, we can use the linear approximation approach (*delta method*). For the predicted probabilities,

$$\text{Asy.Var}[\hat{F}_i] = [\partial \hat{F}_i / \partial \hat{\boldsymbol{\gamma}}]' \, \mathbf{V} \, [\partial \hat{F}_i / \partial \hat{\boldsymbol{\gamma}}],$$

where $\mathbf{V} = \text{Asy.Var}[\hat{\boldsymbol{\gamma}}]$. The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be any of the three described in Section 2.4. Let $z_i = \hat{\boldsymbol{\gamma}}'\mathbf{x}_i$. Then the derivative vector is

$$[\partial \hat{F}_i / \partial \hat{\boldsymbol{\gamma}}] = [d\hat{F}_i / dz][\partial z_i / \partial \hat{\boldsymbol{\gamma}}] = \hat{f}_i \times \mathbf{x}_i.$$

Combining terms gives

$$\text{Asy.Var}[\hat{F}_i] = (\hat{f}_i)^2 \, \mathbf{x}_i'\mathbf{V}\mathbf{x}_i,$$

which depends on the particular $\mathbf{x}_i$ vector used. This result is useful when a marginal effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta \hat{F}_i = [\hat{F}_i \mid (d_i = 1)] - [\hat{F}_i \mid (d_i = 0)].$$

The asymptotic variance would be

$$\text{Asy.Var}[\Delta \hat{F}_i\ ] = [\partial(\Delta \hat{F}_i)/\partial \hat{\boldsymbol{\gamma}}\ ]'\ \mathbf{V}[\partial(\Delta \hat{F}_i)/\partial \hat{\boldsymbol{\gamma}}\ ],$$

where

$$[\partial(\Delta \hat{F}_i)/\partial \hat{\boldsymbol{\gamma}}\ ] = (\hat{f}_i \mid d=1)\begin{pmatrix}\mathbf{x}_{i(d)}\\ 1\end{pmatrix} - (\hat{f}_i \mid d=0)\begin{pmatrix}\mathbf{x}_{i(d)}\\ 0\end{pmatrix}.$$

For the other partial effects,

$$\text{Asy.Var}[\,\hat{\boldsymbol{\delta}}\,] = \left[\frac{\partial \hat{\boldsymbol{\delta}}}{\partial \hat{\boldsymbol{\gamma}}'}\right]\mathbf{V}\left[\frac{\partial \hat{\boldsymbol{\delta}}}{\partial \hat{\boldsymbol{\gamma}}'}\right]'.$$

The matrix of derivatives is

$$\hat{f}\left(\frac{\partial \hat{\boldsymbol{\gamma}}}{\partial \hat{\boldsymbol{\gamma}}'}\right) + \hat{\boldsymbol{\gamma}}\left(\frac{d\hat{f}}{dz}\right)\left(\frac{\partial z}{\partial \hat{\boldsymbol{\gamma}}'}\right) = \hat{f}\ \mathbf{I} + \left(\frac{d\hat{f}}{dz}\right)\hat{\boldsymbol{\gamma}}\bar{\mathbf{x}}'. \tag{2.33}$$

For the probit model, $d\hat{f} / dz = -z\hat{\phi}$, so

$$\begin{aligned}\text{Asy.Var}[\,\hat{\boldsymbol{\delta}}\,] &= \left\{\hat{\phi}[\mathbf{I} - (\hat{\boldsymbol{\gamma}}'\bar{\mathbf{x}})\hat{\boldsymbol{\gamma}}\bar{\mathbf{x}}']\right\}\mathbf{V}\left\{\hat{\phi}[\mathbf{I} - (\hat{\boldsymbol{\gamma}}'\bar{\mathbf{x}})\hat{\boldsymbol{\gamma}}\bar{\mathbf{x}}']\right\}' \\ &= \mathbf{G}_{PROBIT}\,\mathbf{V}\,\mathbf{G}'_{PROBIT}.\end{aligned} \tag{2.34a}$$

For the logit model, $\hat{f} = \hat{\Lambda}(1-\hat{\Lambda})$ so

$$d\hat{f} / dz = (1-2\hat{\Lambda})\left(\frac{d\hat{\Lambda}}{dz}\right) = (1-2\hat{\Lambda})\hat{\Lambda}(1-\hat{\Lambda}).$$

Collecting terms, we obtain

$$\begin{aligned}\text{Asy.Var}[\,\hat{\boldsymbol{\delta}}\,] &= \left\{[\Lambda(1-\Lambda)]\ [\mathbf{I} + (1-2\Lambda)\hat{\boldsymbol{\gamma}}\bar{\mathbf{x}}']\right\}\mathbf{V}\left\{[\Lambda(1-\Lambda)][\mathbf{I} + (1-2\Lambda)\hat{\boldsymbol{\gamma}}\bar{\mathbf{x}}']\right\}'. \\ &= \mathbf{G}_{LOGIT}\,\mathbf{V}\,\mathbf{G}'_{LOGIT}.\end{aligned} \tag{2.34b}$$

As before, the value obtained will depend on the **x** vector used. We have suggested the sample mean above.

2.6.5 Standard Errors for Estimated Odds Ratios

The computation for the estimated odds ratio, $\hat{\tau}_k = \exp(\hat{\gamma}_k)$ is straightforward. Using the delta method, the estimated standard error for $\hat{\tau}_k$ will equal $\hat{\tau}_k$ times the standard error for $\hat{\gamma}_k$. Note, however that the conventional *t* ratio for testing the hypothesis that γ_k equals zero would be inappropriate for τ_k. The appropriate test would be H_0:$\tau_k = 1$, and the t statistic reported should be $t_\tau = (\hat{\tau}_k - 1)$/standard error.

2.6.6 Average Partial Effects

The preceding has emphasized computing the partial effects for the average individual in the sample. Current practice has many applications based, instead, on "average partial effects." [See, e.g.,Wooldridge (2002b).] The underlying logic is that the quantity of interest is

$$APE = E_x\left[\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}}\right].$$

In practical terms, this suggests the computation

$$\begin{aligned}\widehat{APE} &= \bar{\hat{\boldsymbol{\delta}}} \\ &= \frac{1}{n}\sum_{i=1}^{n} f\left(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i\right)\hat{\boldsymbol{\gamma}} \\ &= \left[\frac{1}{n}\sum_{i=1}^{n} f\left(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i\right)\right]\hat{\boldsymbol{\gamma}}.\end{aligned} \tag{2.35}$$

Both the "partial effects at the means" and the "average partial effects" are computed by scaling the coefficient vector. The scale factor for the first is the density evaluated at $\boldsymbol{\gamma}'\bar{\mathbf{x}}$. The scale factor for the second is the sample mean of $f(\boldsymbol{\gamma}'\mathbf{x}_i)$. Table 2.5 shows a comparison of the average partial effects and the partial effects at the sample means for our estimated probit model.

Table 2.5 Marginal Effects and Average Partial Effects

Variable	Marginal Effects	Avg. Partial Effects
AGE	-.00782	-.00751
EDUC	.01999	.01919
INCOME	.08376	.08041
MARRIED	-.01191	-.01146
KIDS	.03066	.02952
Scale Factor	.38417	.36881

This does raise two complications. First, because the computation is (marginally) more burdensome than the simple marginal effects at the means, one might wonder whether this produces a noticeably different answer. That will depend on the data. Generally, except for small sample variation, the difference in these two results is likely to be quite minor, particularly in a large sample. Second, computing the individual effects, then using the natural estimator to estimate the variance of the mean, may badly estimate the asymptotic variance of the average partial effect. [See, e.g. Contoyannis et al. (2004, p. 498).] The natural estimator would be

$$Est.Asy.Var\left[\bar{\hat{\delta}}_k\right] = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\delta}_{i,k} - \bar{\hat{\delta}}_k\right)^2\right].$$

The problem with this computation is that the observations in the *APE* are highly correlated – they all use the same estimate of $\boldsymbol{\gamma}$ – but this variance computation treats them as a random sample. The computations are analyzed in Greene (2008a, pp. 784-785). In principle, the

variance of the mean of n correlated variables should involve n^2 terms (which would be 17 million for our example). But, this computation turns out to be simpler than that; the end result is

$$Est.Asy.Var\left[\bar{\hat{\boldsymbol{\delta}}}\right] = \bar{\mathbf{G}}(\hat{\boldsymbol{\gamma}})\hat{\mathbf{V}}\mathbf{G}(\hat{\boldsymbol{\gamma}})', \tag{2.36}$$

where $\bar{\mathbf{G}}(\hat{\boldsymbol{\gamma}})$ is the sample means of the individual matrices,

$$\bar{\mathbf{G}}(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{\phi}_i[\mathbf{I} - (\hat{\boldsymbol{\gamma}}'\mathbf{x}_i)\hat{\boldsymbol{\gamma}}\mathbf{x}_i']\right\} \text{ for the probit model and} \tag{2.37}$$

$$\bar{\mathbf{G}}(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} \left\{[\hat{\Lambda}_i(1-\hat{\Lambda}_i)][\mathbf{I} + (1-2\hat{\Lambda}_i)\hat{\boldsymbol{\gamma}}\mathbf{x}_i']\right\} \text{ for the logit model.}$$

2.6.7 Standard Errors for Marginal Effects Using the Krinsky and Robb Method

An alternative to the delta method described above that is sometimes advocated is the Krinsky and Robb (1986, 1990, 1991) method. By this device, we have our estimate of the model coefficients, $\hat{\boldsymbol{\gamma}}$, and the estimated asymptotic covariance matrix, $\mathbf{V}$. The marginal effects are computed as a function of $\hat{\boldsymbol{\gamma}}$ and the vector of means of the sample data, $\mathbf{x}$, say $g_k(\hat{\boldsymbol{\gamma}}, \mathbf{x})$ for the kth variable. The Krinsky and Robb technique involves sampling R draws from the asymptotic normal distribution of the estimator, i.e., with mean $\hat{\boldsymbol{\gamma}}$ and covariance $\mathbf{V}$, computing the function with these R draws, then computing the empirical variance. Draws from the required population are obtained as follow: Let $\mathbf{LL}'$ denote the Cholesky factorization of $\mathbf{V}$; $\mathbf{L}$ is a lower triangular matrix. Let $\mathbf{w}_r$ denote the rth vector of K independent draws from the standard normal. Then, $\hat{\boldsymbol{\gamma}}_r = \hat{\boldsymbol{\gamma}} + \mathbf{L}\mathbf{w}_r$. The routine below uses the Krinsky and Robb method to recompute the standard errors for the partial effects that are computed using the delta method in Table 2.4. The simulation uses 1,000 draws. The comparison is shown below. The results with the two methods are nearly identical.

```
Probit   ; lhs=healthy;rhs=x; marginal effects $
Namelist ; x = one,age,educ,income,married,kids $
Matrix   ; xb=mean(x) ; xbc=xb(1:4) $
Calc     ; xb5=xb(5);xb6=xb(6) ; Ran(123457) $ (Set seed for RNG)
Wald     ; k&r ; pts=1000 ; start=b ; var=varb ; labels = a,b1,b2,b3,b4,b5
         ; fn1 = n01(a'xb)*b1  ; fn2 = n01(a'xb)*b2  ; fn3 = n01(a'xb)*b3
         ; fn4 = phi(a'xbc+b4*1+b5*xb6) - phi(a'xbc+b4*0+b5*xb6)
         ; fn5 = phi(a'xbc+b4*xb5+b5*1) - phi(a'xbc+b4*xb5+b5*0) $
```

```
+--------+--------+-------+
|        | St. Er.| St.Er.|
|Variable| Delta  |  K&R  |
+--------+--------+-------+
|AGE     |  .0007 | .0007 |
|EDUC    |  .0034 | .0032 |
|INCOME  |  .0486 | .0490 |
|MARRIED*|  .0194 | .0198 |
|KIDS*   |  .0177 | .0179 |
+--------+--------+-------+
```

2.6.8 Fitted Probabilities

A useful display when the model contains both continuous and interesting discrete variables is a plot of the fitted probabilities that holds the other variables fixed, say at their means,

while simultaneously varying the continuous and discrete variables. This type of plot can show graphically the information contained in the partial effects. For example, Figure 2.5 shows the effect of changes in *Income* on the probability that *Healthy* equals one for 21 year olds and 45 year olds.

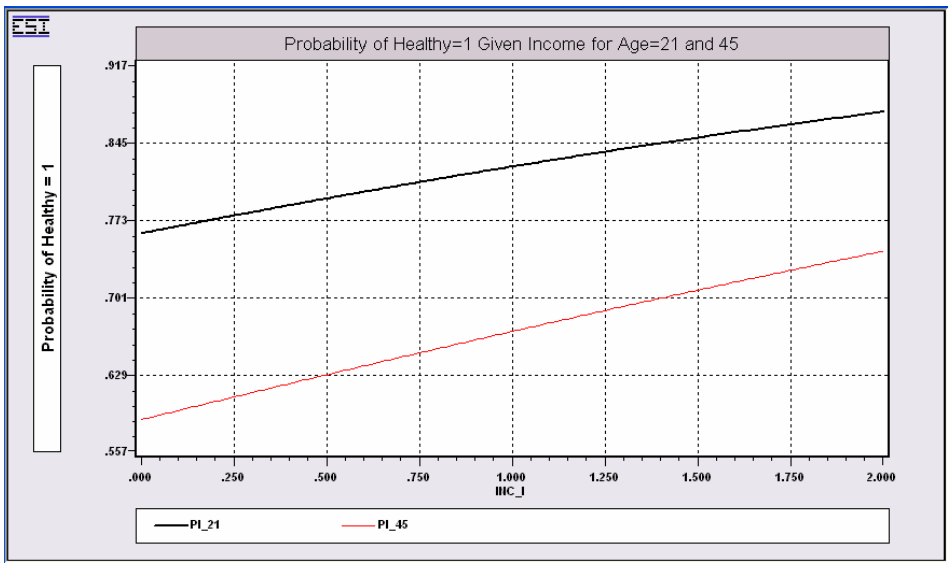


Figure 2.5 Fitted Probabilities for a Probit Model

2.7 Hypothesis Testing

The full standard menu of procedures is available for testing hypotheses about the coefficients. The simplest method for a single restriction would be based on the usual *t* tests, using the standard errors from the estimated asymptotic covariance matrix for the coefficients. Using the normal distribution of the estimator, we would use the standard normal table rather than the *t* table for critical points. Thus, as in conventional regression analysis, the test statistic for testing the null hypothesis that a coefficient equals a specific value,

$$H_0: \gamma_k = \gamma_k^0,$$

would be

$$z = \frac{\hat{\gamma}_k - \gamma_k^0}{\sqrt{Est.Asy.Var(\hat{\gamma}_k)}}.$$

Critical values would be based on the normal distribution, since the distribution used for the statistic holds only asymptotically. The statistic for testing the hypothesis that each coefficient equals zero will be presented with the coefficient estimates by all standard software. See, for example, the third column of results for each model in Table 2.2.

2.7.1 Wald Tests

For more involved restrictions, it is possible to use the Wald test. For a set of restrictions $\mathbf{R}\boldsymbol{\gamma} = \mathbf{q}$, the statistic is

$$W = (\mathbf{R}\hat{\boldsymbol{\gamma}} - \mathbf{q})'[\mathbf{R}\ Est.Asy.Var(\hat{\boldsymbol{\gamma}})\ \mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\gamma}} - \mathbf{q}).$$

This statistic has a chi squared distribution with degrees of freedom equal to the number of restrictions being tested. (That will be the number of rows in **R**.) For example, for testing the hypothesis that a subset of the coefficients, say, the last *M*, are zero, the Wald statistic uses **R** = [**0** | $\mathbf{I}_M$] and **q** = **0**. Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\boldsymbol{\gamma}}_M' \mathbf{V}_{MM}^{-1} \boldsymbol{\gamma}_M ,$$

where the subscript M indicates the subvector or submatrix corresponding to the *M* variables and **V** is the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\gamma}}$. For a set of nonlinear restrictions of the form **r**($\boldsymbol{\gamma}$,**q**) = **0**, based on the delta method, we would use

$$\mathbf{R}\left(\hat{\boldsymbol{\gamma}}\right) = \frac{\partial \mathbf{r}(\hat{\boldsymbol{\gamma}}, \mathbf{q})}{\partial \hat{\boldsymbol{\gamma}}'}$$

in the expression for the Wald statistic. In the linear case, **r**($\boldsymbol{\gamma}$,**q**) = **R**$\boldsymbol{\gamma}$ - **q**.

2.7.2 Likelihood Ratio Tests.

Likelihood ratio and Lagrange multiplier statistics can also be computed. The likelihood ratio statistic is

$$LR = -2[\ln L_R - \ln L_U] , \tag{2.39}$$

where lnL_R and lnL_U are the log-likelihood functions evaluated at the restricted and unrestricted estimates, respectively. The statistic has a limiting chi squared distribution with degrees of freedom equal to the number of restrictions being tested.

A common test, which is similar to the *F* test that all the slopes in a regression are zero, is the *likelihood ratio test* that all the slope coefficients in the probit or logit model are zero. For this test, the constant term remains unrestricted. In this case, the restricted log-likelihood is the same for both probit and logit models,

$$\ln L_0 = n[P_1 \ln P_1 + (1 - P_1) \ln(1 - P_1)], \tag{2.40}$$

where P_1 is the proportion of the observations that have dependent variable equal to 1. The *model chi squared* often reported in statistical results is

$$\chi_{Model}^2 = 2(\ln L - \ln L_0).$$

This is a counterpart to the *F* statistic typically computed for a linear regression model. The statistic is used to test the joint hypothesis that the *K*-1 coefficients on the non-constant variables in the model are all zero.

It might be tempting to use the likelihood ratio test to choose between the probit and logit models. But there is no restriction involved, and the test is not valid for this purpose. To underscore the point, there is nothing in its construction to prevent the chi-squared statistic for this "test" from being negative.

2.7.3 Lagrange Multiplier Tests

The *Lagrange multiplier test* statistic is

$LM = \mathbf{g}'\mathbf{V}\mathbf{g}$,

where **g** is the vector of first derivatives of the *unrestricted* model evaluated at the *restricted* parameter vector and **V** is any of the three estimators of the asymptotic covariance matrix of the maximum likelihood estimator of $\hat{\boldsymbol{\gamma}}$, once again computed using the restricted estimates. A convenient formulation that requires only the first derivatives of the log likelihood is based on the BHHH estimator. For either the probit or logit models the first derivative vector can be written as

$$\frac{\partial \ln L}{\partial \hat{\boldsymbol{\gamma}}} = \sum_{i=1}^{n} \hat{g}_i \mathbf{x}_i ,$$

where $\hat{g}_i$ equals $\left(y_i - \hat{\Lambda}_i\right)$ for the logit model and $\hat{\lambda}_i$ for the probit model. [See (2.14) and (2.17).] Define a diagonal matrix $\hat{\mathbf{G}}$ = diag[$\hat{g}_i$] and let **i** denote an $n\times 1$ column vector of ones. Then,

$$\frac{\partial \ln L}{\partial \hat{\boldsymbol{\gamma}}} = \mathbf{X}'\hat{\mathbf{G}}\mathbf{i} .$$

The BHHH estimator of the Hessian will be $\mathbf{X}'\hat{\mathbf{G}}'\hat{\mathbf{G}}\,\mathbf{X}$, so the LM statistic based on this estimator is

$$LM = n\left[\frac{1}{n}\mathbf{i}'(\hat{\mathbf{G}}\mathbf{X})(\mathbf{X}'\hat{\mathbf{G}}'\hat{\mathbf{G}}\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{G}}')\mathbf{i}\right].$$

Another way to write the statistic which suggests how to set it up for computer programs is

$$LM = \left(\sum_{i=1}^{n} \hat{g}_{i,R}\mathbf{x}_i\right)'\left[\sum_{i=1}^{n} \hat{g}_{i,R}^2 \mathbf{x}_i\mathbf{x}_i'\right]\left(\sum_{i=1}^{n} \hat{g}_{i,R}\mathbf{x}_i\right). \tag{2.41}$$

All the statistics listed here are asymptotically equivalent and under the null hypothesis of the restricted model have limiting chi-squared distributions with degrees of freedom equal to the number of restrictions being tested.

2.7.4 Application of Hypothesis Tests

The application below shows three tests:

a. The individual tests that coefficients equal zero
b. The three tests of the hypothesis that all coefficients except the constant term are zero
c. A homogeneity test. When the sample can be divided into *G* groups, a test of the hypothesis that the same parameter vector applies to all *G* groups is carried out by estimating the model *G*+1 times, once with each group, obtaining log likelihood functions $\ln L_g$ and once with the full pooled data set, obtaining $\ln L_{\text{Pooled}}$. The likelihood ratio test statistic is

$$LR = -2\left[\ln L_{Pooled} - \Sigma_{g=1}^{G} \ln L_g\right]. \quad (2.42)$$

The chi squared statistic has $(G\text{-}1)K$ degrees of freedom, where K is the number of parameters in the model, including the constant term.

Table 2.6 Presents the estimated probit and logit models from Table 2.2. The individual significance tests for the coefficients appear in the column labeled "t." For both models, *Age* and *Educ* are statistically significant while the other variables are not. We would not expect the two models to produce different conclusions. We have also reported the chi squared tests that all coefficients save the constant term are zero. As might be expected, the hypothesis is rejected. (The critical chi squared for 5 degrees of freedom is 11.07 while the model value is about 240. The probit results also include the Wald and LM statistic for the same hypothesis. The similarity to the likelihood ratio statistic is to be expected. Finally, the test for pooling the 7 years of data is carried out in table 2.7. The log likelihood for the pooled sample is -17365.76. The sum of the log likelihoods for the seven individual years is -17324.33. Twice the difference is 82.87. The degrees of freedom is 6×6 = 36. The 95% critical value from the chi squared table is 50.998, so the pooling hypothesis is rejected.

2.8 Goodness of Fit Measures

There have been many fit measures devised for binary choice models. At a minimum, one should report the maximized value of the log-likelihood function, ln$L$. Because the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term, lnL_0, should also be reported. One of the most often reported computations is McFadden's (1974) *likelihood ratio index* or *Pseudo* R^2,

$$Pseudo\ R^2 = LRI = 1 - (\ln L / \ln L_0).$$

Table 2.6 Hypothesis Tests

	Logit				Probit				
	LogL = -2890.393				LogL = -2890.288				
	LogL0 = -3010.421				LogL0 = -3010.421				
	Chisq = 240.056				Chisq = 240.266				
					Wald = 234.349				
					LM = 238.677				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.7595	.2349	3.233	.0012	.4816	.1423	3.383	.0007	1.0000
AGE	-.0329	.0032	-10.266	.0000	-.0203	.0020	-10.386	.0000	43.4401
EDUC	.0860	.0148	5.805	.0000	.0520	.0089	5.872	.0000	11.4181
INCOME	.3454	.2083	1.658	.0972	.2180	.1265	1.724	.0847	.34874
MARRIED	-.0483	.0828	-.584	.5592	-.0311	.0508	-.612	.5403	.75217
KIDS	.1278	.0756	1.692	.0907	.0800	.0463	1.727	.0841	.37943

Table 2.7 Homogeneity Test

Year	Log Likelihood Function	Sample Size
1984	-2395.137	3874
1985	-2375.090	3794
1986	-2387.602	3792
1987	-2337.835	3666
1988	-2890.288	4483
1991	-2769.375	4340
1994	-2168.998	3377
Pool	-17365.76	27326

This measure has an intuitive appeal in that it is bounded by zero and one. If all the slope coefficients are zero, then it equals zero. There is no way to make *LRI* equal one, although one can come close. If F_i is always one when y_i equals one and always zero when y_i equals zero, then ln L equals zero (the log of one) and *LRI* equals one. It has been suggested that this finding is indicative of a "perfect fit" and that *LRI* increases as the fit of the model improves. To a degree, this point is true. Unfortunately, the values between zero and one have no natural interpretation. If $F(\boldsymbol{\gamma}'\mathbf{x}_i)$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\boldsymbol{\gamma}'\mathbf{x}_i$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. [See, for example, Cragg and Uhler (1970), Amemiya (1981), Maddala (1983), McFadden (1974), Ben-Akiva and Lerman (1985), Kay and Little (1986), Veall and Zimmermann (1992), Zavoina and McKelvey (1975), Efron (1978), and Cramer (1999). A survey of techniques appears in Windmeijer (1995).]

It should be emphasized that whatever its value, the Pseudo R^2 has no connection to a "proportion of variation explained." The dependent variable in the model is only zeros and ones, and the "variation" such as it is, has not appeared in any of the computations we have done. In this regard, it is not, in fact, an analog to R^2 in regression. One other point worth noting is that the *LRI* should never be computed for any model that is not a discrete choice model. The reason is that it is only in discrete choice models that the log likelihood is guaranteed to be negative. When the dependent variable is continuous, for example in linear regression, the log likelihood function can be positive or negative, and *LRI* can take any value.

The Akaike (1973) information criterion (AIC) statistic or its log,

$$\ln AIC = (-2\ln L + 2K)/n,$$

is a fit measure based on the likelihood function that is like the adjusted R^2 in linear regression in that it "rewards" good fit but penalizes the model for having a large number of parameters. The *AIC* measure is often used to compare nonnested models when there is no obvious criterion or rule for comparing fits. There is no distribution theory for AIC or lnAIC that produces a formal test of any hypothesis. Rather, the statistic is used as a practical measure for comparing models, for example, in cases in which the models are nonnested.

Some authors have proposed other "fit measures" that are based on the log likelihood function. Veall and Zimmermann's (1992) suggested measure is

$$R^2_{VZ} = \left(\frac{\delta - 1}{\delta - LRI}\right) LRI,\ \delta = \frac{n}{2\ln L_0}.$$

Another is

$$R^2_{ML} = 1 - \exp\left[\frac{2(\ln L_0 - \ln L)}{n}\right] = 1 - \exp\left[\frac{-\chi_0^2}{n}\right],$$

where χ_0^2 is the likelihood ratio statistic used to test the hypothesis that all the coefficients in the model are zero. Like the LRI, in spite of their names, these are not fit measures as such, nor, for that matter are they correlation coefficients. These are all function of the log likelihood function that are bounded by zero and one and that increase, albeit at different rates, when variables are added to the model.

2.8.1 Perfect Prediction

If the range of one of the independent variables contains a value, say, x^*, such that the sign of $(x - x^*)$ predicts y perfectly and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of $\boldsymbol{\gamma}'\mathbf{x}_i$ gives a perfect predictor for some vector $\boldsymbol{\gamma}$. For example, one might mistakenly include as a regressor a dummy variables that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because $\boldsymbol{\gamma}'\mathbf{x}_i$ is diverging during the iterations. [See McKenzie (1998) for an application and discussion.] Of course, this situation is not at all what we had in mind for a good fit.

2.8.2 Dummy Variables with Empty Cells

A problem similar to the one noted above arises when a model includes a dummy variable that has no observations in one of the cells of the dependent variable. An example appears on page 673 of the fifth (1993) edition of Greene (2003), in which the dependent variable is always zero when the variable 'Southwest' is zero. McKenzie (1998) and Stokes (2004) have used this example and others to examine a number of econometrics programs. He found that no program which did not specifically check for the failure – only one did – could detect the failure in some other way. All iterated to apparent convergence, though with very different estimates of this coefficient and differing numbers of iterations because of their use of different convergence rules. This form of incomplete matching of values likewise prevents estimation, though the effect is likely to be more subtle. In this case, a likely outcome is that the iterations will fail to converge, though the parameter estimates will not necessarily become extreme.

2.8.3 Explaining Variation in the Implied Regression

The fit measures suggested do not actually correspond to the conventional measure of fit in a regression model, that is, the ability of the model to predict the dependent variable. One might interpret the model directly as a nonlinear regression, since

$$\begin{aligned} E[y_i|\underline{\mathbf{x}}_i] &= 0 \times (1 - \mathrm{F}_i) + 1 \times F_i \\ &= F_i. \end{aligned}$$

It follows that

$$\begin{aligned} u_i &= y_i - F\left(\hat{\boldsymbol{\gamma}}'\mathbf{x}\right) \\ &= y_i - \hat{F}_i \end{aligned}$$

is a bona fide residual, albeit one that is always equal either to $\hat{F}_i$ or 1 - $\hat{F}_i$. With this in hand, Efron's (1978) proposed fit measure is

$$R^2_{Ef} = 1 - \frac{\Sigma_{i=1}^n u_i^2}{\Sigma_{i=1}^n (y_i - \overline{y})^2}$$
$$= 1 - \frac{\frac{1}{n}\Sigma_{i=1}^n u_i^2}{P_1(1-P_1)}.$$

The statistic bears some resemblance to the conventional R^2 in regression if one considers the denominator the total variation of the binary dependent variable. If the model contains only a constant term, then R^2_{Ef} equals zero because $\hat{F}_i$ will equal $\overline{y}$ for every observation. How the statistic behaves when one adds variables to an existing model is uncertain, however. Nothing in the construction guarantees that the numerator in the fraction will fall when variables are added to a model that already contains at least one variable. Also, it should not be interpreted as a "proportion of variation explained," as it is not bound in the [0,1] interval.

Notwithstanding its somewhat ambiguous nature, a modification of u_i that has been suggested is the *Pearson residual*, [see Johnson and Albert (1999, p.94)],

$$u_{i,P} = \frac{u_i}{\hat{F}_i\left(1-\hat{F}_i\right)}.$$

(The authors' suggestion that the distribution of $u_{i,P}$ can be reasonably approximated by a standard normal distribution is clearly inappropriate for binary data, though it might work better for grouped date when $y_i = p_i$, the proportion of n_i observations with identical $\mathbf{x}_i$ that "respond" with outcome equal to 1, such as in bioassay or in a clinical trial.) Two suggested refinements on this computation (for binary data) are the *deviance residuals*,

$$u_{i,D} = q_i\sqrt{-2\ln L_i},$$

where $q_i = 2y_i - 1$ and L_i is the contribution of observation i to the likelihood function and the *adjusted deviance residuals*,

$$u_{i,AD} = u_{i,D} + \frac{1-2\hat{F}_i}{6\sqrt{\hat{F}_i\left(1-\hat{F}_i\right)}}.$$

The authors suggest that plots of the residuals against the fitted probabilities can be helpful in identifying outliers in the data.

McKelvey and Zavoina's (1975) suggestion is based on the latent regression,

$$R^2_{MZ} = \frac{\Sigma_{i=1}^n \left(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i - \hat{\boldsymbol{\gamma}}'\overline{\mathbf{x}}\right)}{n + \Sigma_{i=1}^n \left(\hat{\boldsymbol{\gamma}}'\mathbf{x}_i - \hat{\boldsymbol{\gamma}}'\overline{\mathbf{x}}\right)}.$$

The McElvey and Zavoina measure corresponds to the ratio of the regression variation to the total variation in the latent regression $y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i$. The computation is made possible because the

variance of ε_i is known to equal one for a probit model. Note, as well, that this computation will differ substantively when used for a logit model, since the sums will be multiplied by roughly 1.6^2 while the n×1 in the denominator must be replaced with $n\pi^2/3$. It follows that R_{MZ}^2 will be systematically lower for a logit model than a probit model. Since prediction of y_i* is rarely the objective of estimation, this measure is not commonly used.

2.8.4 Fit Measures Based on Predicted Probabilities

A fundamental flaw in the fit measures already suggested is that although they are labeled R^2 measures, with the exception of the Efron measure, they do not, in fact, measure the fit of the model to the data. The likelihood function is not maximized so as to minimize the distance between F_i and y_i. Other fit measures have been suggested that are more in line with this objective. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

$$R_{BL}^2 = \frac{1}{n}\sum_{i=1}^{n} \left[y_i \hat{F}_i + (1-y_i)(1-\hat{F}_i) \right].$$

This is the average probability of correct prediction of y_i using $\hat{F}_i$. The difficulty in this computation is that in unbalanced samples, the less frequent outcome will usually be predicted very badly by the standard procedure, and this measure does not pick up that point. Cramer (1999) has suggested an alternative measure that directly measures this failure,

$$\begin{aligned}\lambda &= (\text{average } \hat{F}_i \mid y_i = 1) - (\text{average } \hat{F}_i \mid y_i = 0) \\ &= (\text{average } (1\text{-}\hat{F}_i) \mid y_i = 0) - (\text{average } (1\text{-}\hat{F}_i) \mid y_i = 1).\end{aligned}$$

Cramer's measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes.

Table 2.8 reports the various fit measures for the probit model. The small values are somewhat surprising, given the results in the next section that show the model actually does quite a good job in predicting the outcome variable. The very large difference between the Ben Akiva/Lerman measure and Cramer's statistic underscores the need to look carefully at these results when reporting them.

Table 2.8 Fit Measures for Probit Model

Proportions P0= .396386	P1= .603614	
N = 4483	N0= 1777	N1= 2706
LogL= -2890.288	LogL0= -3010.421	
Efron	McFadden	Ben./Lerman
.05254	.03991	.54668
Rsqrd_ML	Veall/Zim.	Cramer
.05218	.08874	.05262
Akaike InformationCriterion 1.29212		

2.8.5 Assessing the Model's Ability to Predict

A useful summary of the predictive ability of the model is a 2 × 2 table of the hits and misses of a prediction rule such as

$$\hat{y} = 1 \text{ if } \hat{F}_i > F^* \text{ and 0 otherwise.}$$

The usual threshold value is 0.5, on the basis that we should predict a one if the model says a one is more likely than a zero. Table 2.9 shows the results for our probit model for Healthy. Although the R^2 measures based on the log likelihood function are all very small, less than 0.1, the model correctly predicts 62% of the observations. This demonstrates the substantial disconnecton between these two notions of "fit."

Table 2.9 Prediction Success for Probit Model

Predictions for Binary Choice Model. Predicted value is 1 when probability is greater than .500000, 0 otherwise.			
Actual Value	Predicted Value 0	Predicted Value 1	Total Actual
0	530 (11.8%)	1247 (27.8%)	1777 (39.6%)
1	456 (10.2%)	2250 (50.2%)	2706 (60.4%)
Total	986 (22.0%)	3497 (78.0%)	4483 (100.0%)

A number of summary measures can be constructed for the success of the model to predict the outcome using this rule. A list is shown in Table 2.10 for the health care example.

Table 2.10 Success Measures for Predictions by Estimated Probit Model

Analysis of Binary Choice Model Predictions Based on Threshold = .5000	
Prediction Success	
Sensitivity = actual 1s correctly predicted	83.149%
Specificity = actual 0s correctly predicted	29.826%
Positive predictive value = predicted 1s that were actual 1s	64.341%
Negative predictive value = predicted 0s that were actual 0s	53.753%
Correct prediction = actual 1s and 0s correctly predicted	62.012%
Prediction Failure	
False pos. for true neg. = actual 0s predicted as 1s	70.174%
False neg. for true pos. = actual 1s predicted as 0s	16.851%
False pos. for predicted pos. = predicted 1s actual 0s	35.659%
False neg. for predicted neg. = predicted 0s actual 1s	46.247%
False predictions = actual 1s and 0s incorrectly predicted	37.988%

These fit measures can be problematic in highly unbalanced samples, that is, that have many more ones than zeros, or vice versa. Consider, for example, the naive predictor, always predict $\hat{y} = 1$ if $P > 0.5$ and 0 otherwise, where P is the simple proportion of ones in the sample. This rule will always predict correctly $100P$ percent of the observations, which means that the naive model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the naive predictor (no model) will generate more correct predictions than the prediction rule with a fuller model. Once again, this flaw is not in the model; it is a flaw in the fit measure. The important element to bear in mind is that the

coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where **b** maximizes R^2. Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is heavily unbalanced, then this prediction rule might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1,000 have $y_i = 1$.We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce an $\hat{F}_i$ of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $y_i = 1$. The obvious adjustment is to reduce F^*. Of course, this adjustment comes at a cost. If we reduce the threshold F^* so as to predict $y_i = 1$ more often, then we will increase the number of correct classifications of observations that do have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y_i = 0$. In general, any prediction rule of the form $\hat{y} = 1$ if $\hat{F}_i > F^*$ and 0 otherwise. will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model [see Boyes, Hoffman, and Low (1989)], incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one. Changing F^* will always reduce the probability of one type of error while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends. Figure 2.6 shows the tradeoff inherent in choosing different thresholds for the health care example.

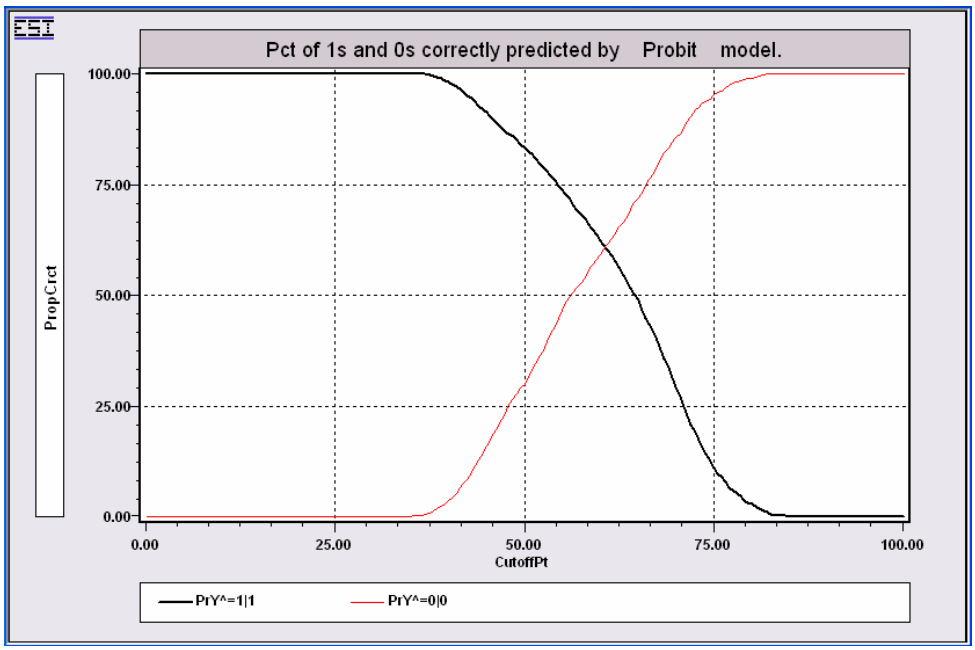


Figure 2.6 Prediction Success for Different Prediction Rules

2.8.6 A Specification Test Based on Fit

Hosmer and Lemeshow (2000) have proposed a diagnostic measure for the probit and logit models (they focus on the latter) that assesses the match between actual and predicted values. To do the computation, we compute a fitted probability, F_i for each observation using the estimated model parameters. We then sort the fitted values in ascending order, carrying the actual y_i with them. The data are then divided into 10

percentiles based on the fitted values, and means of the predicted and actual data are computed within each group. The statistic is

$$H = \sum_{j=1}^{10} n_j \left[\frac{\left(\bar{y}_j - \bar{\hat{F}}_j \right)^2}{\bar{\hat{F}}_j \left(1 - \bar{\hat{F}}_j \right)} \right].$$

(If the sample is not large, some groups at the high or low end may have insufficient variation to compute the denominator – the fitted values may all be very close to zero or one. The resulting statistic has a limiting chi squared distribution with 10 degrees of freedom. Large values of the statistic suggest that the model is inappropriate. The example for the health care data below suggests this case. The H statistic for the model in Table 2.2 is 16.789 with 8 degrees of freedom. The P value is 0.03238 which casts doubt on the distributional assumption.

2.8.7 ROC Plots for Binary Choice Models

Receiver operating characteristic (ROC) plots provide a loose descriptive measure of fit in a binary choice model, and can be used to some extent to compare models. An example appears in Figure 2.7. The curve is constructed by computing for the range of values of P^* from zero to one, the Sensitivity(P^*) which equals the proportion of observations for which estimated and actual values of y_i are both equal to one (when the estimated y_i equals one if the the predicted probability is greater than or equal to P^*). The Specificity(P^*) equals the proportion of values for which predicted and actual zeros match. The graph is constructed by plotting Sensitivity(P^*) against 1 - Specificity(P^*). The 'fit measure' is then computed as the area under the ROC curve. A greater area implies a greater model fit. (The field is a unit rectangle.) A model with no fit has an area of 0.5.

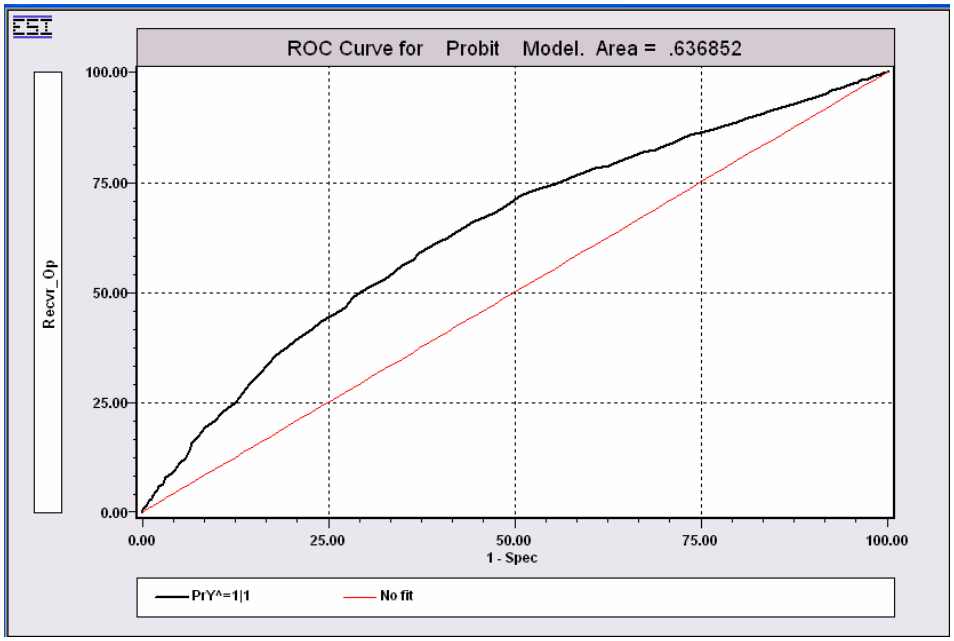


Figure 2.7 ROC Curve for Estimated Probit Model

2.9 Heteroscedasticity

The assumption of homoscedasticity of ε_i in the binary choice model (and the ordered choice model discussed later), is likely to be violated in micro- level data. Unfortunately, there are no robust parametric approaches to model fitting and analysis. Moreover, semiparametric approaches that are robust to heteroscedasticity such as maximum score [Manski (1995)] and the Klein and Spady (1993) approach, have a fundamental shortcoming; only ratios of partial effects and coefficients can be estimated and fit measures, to the extent that they measure anything, are meaningless in these contexts. For better or worse, formal treatment of heteroscedasticity in binary choice models must be specified parametrically in terms of observables. (I.e., there is no counterpart to the White (1980) estimator for unspecified heteroscedasticity.)

We use the general formulation analyzed by Harvey (1976),

$$\text{Var}[\varepsilon_i|\mathbf{z}_i] = [\exp(\boldsymbol{\theta}'\mathbf{z}_i)]^2.$$

This model can be applied equally to the probit and logit models. We will derive the results specifically for the probit model; the logit model is essentially the same. Thus,

$$y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \varepsilon_i,$$
$$\text{Var}[\varepsilon_i \mid \mathbf{x},\mathbf{z}] = [\exp(\boldsymbol{\theta}'\mathbf{z}_i)]^2.$$

The presence of heteroscedasticity necessitates some care in interpreting the coefficients for a variable w_k that could be in $\mathbf{x}$ or $\mathbf{z}$ or both. The partial effects are

$$\frac{\partial \text{Prob}(y_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i)}{\partial w_{ik}} = \phi\left[\frac{\boldsymbol{\gamma}'\mathbf{x}_i}{\exp(\boldsymbol{\theta}'\mathbf{z}_i)}\right]\frac{\gamma_k - (\boldsymbol{\gamma}'\mathbf{x}_i)\theta_k}{\exp(\boldsymbol{\theta}'\mathbf{z}_i)}. \tag{2.43}$$

Only the first (second) term applies if w_k appears only in $\mathbf{x}$ ($\mathbf{z}$). This implies that the simple coefficient may differ radically from the effect that is of interest in the estimated model. [See Knapp and Seaks (1992).] The log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \ln F\left[q_i \frac{\boldsymbol{\gamma}'\mathbf{x}_i}{\exp(\boldsymbol{\theta}'\mathbf{z}_i)}\right]. \tag{2.44}$$

To be able to estimate all the parameters, $\mathbf{z}_i$ cannot include a constant term. The derivatives are

$$\frac{\partial \ln L}{\partial\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\theta}\end{pmatrix}} = \sum_{i=1}^{n}\left[\exp(-\boldsymbol{\theta}'\mathbf{z}_i)\frac{f_i(y_i - F_i)}{F_i(1-F_i)}\right]\begin{pmatrix}\mathbf{x}_i\\ \mathbf{z}_i(-\boldsymbol{\gamma}'\mathbf{x}_i)\end{pmatrix}, \tag{2.45}$$

which implies a difficult log-likelihood to maximize – though the model is provided as a built in procedure in *NLOGIT* and *Stata*.

The LM test provides a convenient way to test for heteroscedasticity. The model is easily estimated assuming that $\boldsymbol{\theta} = \mathbf{0}$, as this is the probit or logit model we began with. Let $\mathbf{w}_i$ equal the data vector in parentheses in the derivatives of the log likelihood in (2.45) and let g_i be the term in square brackets. Then, the *LM* statistic is just

$$LM = \left(\sum_{i=1}^{n} g_i \mathbf{w}_i\right)' \left[\sum_{i=1}^{n} g_i^2 \mathbf{w}_i \mathbf{w}_i'\right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{w}_i\right).$$

The likelihood ratio or Wald statistics are also straightforward to compute if one is able to estimate the unrestricted heteroscedastic model.

An application is shown in Table 2.11. We have modeled the variance function in terms of *Income*, *Kids*, *Female* and *Working*, a dummy variable for whether the respondent is employed at the time of the survey. The results carry out the LR, Wald and Lagrange multipllier tests of homoscedasticity. The coefficients in the variance function are constrained to zero. The LM statistic is 3.8577 with two degrees of freedom. The critical value (95%) is 5.99, so the hypothesis of homoscedasticity is not rejected. The second set of results are for the model with heteroscedasticity. The likelihood ratio statisic is LR=2[(-2888.328) - (-2890.288)] = 3.92. The conclusion is the same. The Wald test based on the unrestricted (heteroscedastic) model is 3.72828, leading to the same inference. The coefficient estimates are shown in the table as well. Overall, the data do not suggest that there is heteroscedasticity present. Partial effects for the restricted and unrestriced models are shown at the end of Table 2.11. The index function and variance function have two variables in common, *Income* and *Kids*. Partial effects are computed as the sum of the two terms shown in (2.43). The change from the homoscedastic model is minor for this model.

Table 2.11 Heteroscedastic Probit Model

	Heteroscedastic				Homoscedastic				
	LogL = -2888.328				LogL = -2890.288				
	LogLR = -2890.288				LogL0 = -3010.421				
	Chisq = 3.920				Chisq = 240.266				
	Wald = 3.728								
	LM = 3.858								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.7595	.2349	3.233	.0012	.4816	.1423	3.383	.0007	1.0000
AGE	-.0329	.0032	-10.266	.0000	-.0203	.0020	-10.386	.0000	43.4401
EDUC	.0860	.0148	5.805	.0000	.0520	.0089	5.872	.0000	11.4181
INCOME	.3454	.2083	1.658	.0972	.2180	.1265	1.724	.0847	.34874
MARRIED	-.0483	.0828	-.584	.5592	-.0311	.0508	-.612	.5403	.75217
KIDS	.1278	.0756	1.692	.0907	.0800	.0463	1.727	.0841	.37943
Variance Function									
INCOME	.0141	.5193	.027	.9784					.34874
KIDS	-.1608	.1975	-.814	.4158					.37943
FEMALE	.0291	.1073	.271	.7864					.48405
WORKING	-.1831	.1350	-1.356	.1750					.67232
Partial Effects					Partial Effects				
AGE	-.0080	.0008	-9.469	.0000	-.0078	.0008	-10.392	.0000	43.4401
EDUC	.0190	.0035	5.443	.0000	.0200	.0034	5.875	.0000	11.4181
INCOME	.0859	.1539	.558	.5769	.0838	.0486	1.724	.0847	.34874
MARRIED	-.0171	.0217	-.789	.4301	-.0119	.0194	-.614	.5394	.75217
KIDS	.0314	.0478	.657	.5113	.0307	.0177	1.733	.0831	.37943
FEMALE	-.0029	.0104	-.282	.7779					.48405
WORKING	.0184	.0186	.989	.3227					.67232

2.10 Panel Data

A structural model for a possibly unbalanced panel of data would be written

$$y_{it}^* = \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it},\ i = 1,...,n,\ t = 1,...,T_i,$$

$$y_{it} = 1 \text{ if } y_{it}^* > 0 \text{ and 0 otherwise.}$$

Ideally, we would like to specify that ε_{it} and ε_{is} are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing joint probabilities from a T_i variate distribution, which is generally problematic. (See Section 2.14.) A more limited approach is an *effects model*,

$$y_{it}^* = \boldsymbol{\gamma}'\mathbf{x}_{it} + v_{it} + u_i,\ i = 1,...,n,\ t = 1,...,T_i,$$
$$y_{it} = 1 \text{ if } y_{it}^* > 0 \text{ and 0 otherwise.}$$

where u_i is the unobserved, individual specific heterogeneity. We distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the *random effects model***.** Note that this places a restriction on the distribution of the heterogeneity. If that distribution is unrestricted, so that u_i and $\mathbf{x}_{it}$ may be correlated, then we have the *fixed effects model***.** The distinction does not relate to any intrinsic characteristic of the effect, itself. This is a modeling framework that is fraught with difficulties and unconventional estimation problems. Prominent among them are the following:

> Estimation of the random effects model requires very strong assumptions about the heterogeneity.
>
> The fixed effects model encounters an *incidental parameters problem* that renders the maximum likelihood estimator inconsistent even when the model is properly specified.
>
> As in the linear model, there cannot be any time invariant variables in a fixed effects binary choice model. The time invariant variables become indistinguishable from the fixed effects.

This is a pessimistic beginning. We will develop an approach that suggests at least a partial path around these shortcomings.

2.10.1 Pooled Estimation, Clustering and Robust Covariance Matrix Estimation

If the appropriate model is either a fixed or random effects specification (or any other specification that involves correlation across observations), then the pooled estimator obtained by ignoring the panel nature of the data will be inconsistent. We will obtain an explicit expression for the random effects case below. Assume, however, that the pooled estimator is consistent for some vector of constants – perhaps even one that is useful. In the same manner that the covariance matrix computed for OLS in a linear model with random effects is inappropriate, the covariance matrix computed for the pooled probit or logit estimator will not estimate the correct asymptotic covariance. A computation that is often used is the *cluster corrrected covariance matrix*. [See Wooldridge (2008).

The pooled MLE based on using Newton's method or a similar algorithm,

$$\hat{\boldsymbol{\gamma}} = \text{plim}\hat{\boldsymbol{\gamma}} + \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{H}_{it}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{g}_{it}\right) + o(1/\Sigma_i T_i), \tag{2.46}$$

where $\mathbf{H}_{it}$ is the contribution of individual it to the second derivatives matrix, $\mathbf{g}_{it}$ is the first derivative vector and the final term is the sampling error that vanishes as n increases – T_i does

not. We have used plim $\hat{\boldsymbol{\gamma}}$ rather than γ in the first term because the estimator is likely to be inconsistent. The result can be written

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &\approx \text{plim}\hat{\boldsymbol{\gamma}} + \left(\frac{1}{\sum_{i=1}^{n} T_i}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{H}_{it}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{g}_{it}\right) \\ &= \text{plim}\hat{\boldsymbol{\gamma}} + \left(\bar{\bar{\mathbf{H}}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{g}}_i\right).\end{aligned} \tag{2.47}$$

Assuming that $\bar{\bar{\mathbf{H}}}$ converges to a finite negative definite matrix, the implied estimator for the asymptotic covariance matrix should be

$$Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}\right] = \left(\bar{\bar{\mathbf{H}}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{g}}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{g}}_i\right)'\left(\bar{\bar{\mathbf{H}}}\right)^{-1}. \tag{2.48}$$

The terms with unequal subscripts in the double sum in the middle term correspond to different individuals. Since observations are independent, these terms should (in aggregate) converge to zero. This would leave

$$Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}\right] = \left(\bar{\bar{\mathbf{H}}}\right)^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\bar{\mathbf{g}}_i\bar{\mathbf{g}}_i'\right)\left(\bar{\bar{\mathbf{H}}}\right)^{-1}. \tag{2.49}$$

This is the *cluster corrected covariance matrix* for the binary choice estimator. A refinement that is sometimes employed [See Stata (2007)] is

$$Est.Asy.Var\left[\hat{\boldsymbol{\gamma}}\right] = \left(\bar{\bar{\mathbf{H}}}\right)^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}(\mathbf{g}_{it}-\bar{\mathbf{g}}_i)(\mathbf{g}_{it}-\bar{\mathbf{g}}_i)'\right)\right)\left(\bar{\bar{\mathbf{H}}}\right)^{-1}. \tag{2.50}$$

The assertion of robustness is dubious. First, the estimator is not robust to any conceivable failure of the assumptions of the model. The pooled MLE will be inconsistent for $\boldsymbol{\gamma}$ regardless of the nature of the correlation across observations. On the other hand, the direct question would be whether the cluster corrected estimator is a robust estimator of the asymptotic covariance matrix for the pooled estimator, regardless of what the estimator, itself converges to. Several assumptions have been made to reach this point, so the answer is uncertain. In practical terms, the estimator usually differs substantively from that based on $\left(\bar{\bar{\mathbf{H}}}\right)^{-1}$, which is at least suggestive that the simple pooled estimator is, itself, not robust to the failure of the pooling assumption. The example in Table 2.12 is illustrative. The estimator uses the full unbalanced panel of 27,326 observations. The corrected and uncorrected standard errors are shown with the estimates. The correction produces a 40% - 50% increase in the standard errors. This is typical for applications in which there is a significant degree of correlation across observations that is ignored by the pooled estimator.

Table 2.12 Cluster Corrected Covariance Matrix (7293 Groups)

Variable	Coefficient	Standard Error Pooled	Standard Error Cluster Cor.
Constant	.49632326	.05891212	.08678277
AGE	-.02317830	.00079949	.00111641
EDUC	.05732077	.00370607	.00578509
INCOME	.34245820	.04810999	.06162735
MARRIED	.01293268	.02062755	.02926480
KIDS	.06657821	.01859187	.02493187

2.10.2 Fixed Effects

The fixed effects model is

$$y_{it}^{*} = \alpha_i d_{it} + \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it},\ i = 1,...,n,\ t = 1,...,T_i,$$
$$y_{it} = 1 \text{ if } y_{it}^{*} > 0 \text{ and 0 otherwise.}$$

where d_{it} is a dummy variable that takes the value one in every period for individual i and zero otherwise. For convenience,we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model. The parameters to be estimated are the K elements of $\boldsymbol{\gamma}$ and the n individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters, $(n + K)$ − the number of groups, n is not limited here, and could be in the thousands in a typical application. The log likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln F\left[q_{it}(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_{it})\right],$$

where $F(.)$ is the probability of the observed outcome, $\Phi\left[q_{it}(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_{it})\right]$ for the probit model or $\Lambda\left[q_{it}(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_{it})\right]$ for the logit model. It will be convenient to let $z_{it} = \alpha_i + \boldsymbol{\gamma}'\mathbf{x}_{it}$ so $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = F(q_{it}z_{it})$.

In the linear regression case, estimation of the parameters is made possible by transforming the data to deviations from group means. This eliminates the individual specific constants from the estimator. That trick will not be usable here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time as $\boldsymbol{\gamma}$. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] A method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Greene (2008a, Section 16.9.6.c). The problems with the fixed effects estimator are statistical, not practical. The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are inconsistent (not because they converge to something other than α_i, but because they do not converge at all). The estimator of $\boldsymbol{\gamma}$ is a function of the estimators of α_i, which means that the MLE of $\boldsymbol{\gamma}$ is not consistent either. This is the *incidental parameters problem*. [See Neyman and Scott (1948) and Lancaster (2000).] There is a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman's (1981a,b) results for the probit

model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\gamma}$ is 100 percent, which is extremely pessimistic. Heckman found in a small Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in *very* few cases, it can be shown that there is no incidental parameters problem. (The Poisson regression model is one of these special cases.) A 100% bias for the probit estimator has been widely observed [e.g., Katz (2001), Greene (2004)], but not proven analytically. The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in Greene (2004, 2008a, Chapter 17).

Estimates of a fixed effects probit model are presented in Table 2.13. The results indicate that 3,289 individuals were dropped from the sample. These are the individuals who had y_{it} equal one or zero in every period. Except for the income coefficient, which is surprisingly stable, the fixed effects estimates and partial effects are quite different from the pooled results. Given the very large change in the log likelihood function, this is not surprising. The likelihood ratio test against the null hypothesis of no effects is over 17,700. The partial effects also change substantially when the effects are added to the model. Since the group sizes are small (T_i ranges from 1 to 7), the slope estimator is inconsistent. Whether this is propogated to the estimates of the partial effects ramains to be established.

Table 2.13 Fixed Effects Probit Model

	Fixed Effects				Pooled				
	LogL = -8500.704				LogL = -17365.76				
	LogLR = -17365.76				LogL0 = -18279.95				
	7293 Individuals								
	3289 Individuals Bypassed								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant					.4963	.0589	8.425	.0000	1.0000
AGE	-.0649	.0045	-14.418	.0000	-.0232	.0008	-28.991	.0000	43.5257
EDUC	.0027	.0506	.054	.9570	.0573	.0037	15.467	.0000	11.3206
INCOME	.3530	.1161	3.040	.0024	.3425	.0481	7.118	.0000	.35208
MARRIED	-.0609	.0666	-.915	.3600	.0129	.0206	.627	.5307	.75862
KIDS	-.0118	.0475	-.249	.8032	.0666	.0186	3.581	.0003	.40273
	Partial Effects				Partial Effects				
AGE	-.0248	.0049	-5.087	.0000	-.0089	.0003	-29.012	.0000	43.5257
EDUC	.0010	.0192	.054	.9567	.0219	.0014	15.478	.0000	11.3206
INCOME	.1349	.0515	2.617	.0089	.1309	.0184	7.118	.0000	.35208
MARRIED	-.0233	.0010	-22.562	.0000	.0049	.0079	.626	.5311	.75862
KIDS	-.0045	.0004	-10.792	.0000	.0254	.0071	3.589	.0003	.40273

The incidental parameters problem in estimation of the slope parameters arises here and (apparently) not in the linear regression model. Estimation in the regression model is based on the deviations from group means, not the original data as it is here. The result exploited there is that although $f(y_{it} \mid \mathbf{X}_i)$ is a function of α_i, $f(y_{it} \mid \mathbf{X}_i, \bar{y}_i)$ is not a function of α_i, and the latter is used in least squares estimation of $\boldsymbol{\gamma}$. In the regression setting, $\bar{y}_i$ is a *sufficient statistic* for α_i. Sufficient statistics are available for a few distributions, but not for the probit model. They are available for the logit model, as we now examine. Before considering the alternative estimator, we note, the absence of the incidental parameters problem in the regression is, in fact, only apparent. The MLE of σ^2 in the fixed effects linear regression model (assuming a balanced panel) is $\mathbf{e}'\mathbf{e}/(nT)$, which converges to $[(T\text{-}1)/T]\sigma^2$. If T is small, σ^2 may be significantly underestimated (e.g., by 50% if $T = 2$). The problem shows up in the scaling parameter, not the

slopes. We might note, implicitly, the probit MLE is estimating $\boldsymbol{\gamma}/\sigma^2$. The case of the +100% bias in the fixed effects probit MLE is perhaps not surprising.

A fixed effects binary logit model is

$$\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it})=\frac{\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{it})}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{it})}.$$

The unconditional likelihood for the nT independent observations is

$$L=\prod_{i=1}^{n}\prod_{t=1}^{T_i}\frac{\exp[q_{it}(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{it})]}{1+\exp[q_{it}(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{it})]}.$$

Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the *conditional likelihood function***,**

$$L_c=\prod_{i=1}^{n}\text{Prob}(Y_{i1}=y_{i1},Y_{i2}=y_{i2},...,Y_{iT_i}=y_{iT_i}\,|\,\mathbf{X}_i,\Sigma_{t=1}^{T_i}y_{it}),$$

is free of the incidental parameters, α_i .The joint likelihood for each set of T_i observations conditioned on the number of ones in the set is

$$\prod_{i=1}^{n}\text{Prob}(Y_{i1}=y_{i1},Y_{i2}=y_{i2},...,Y_{iT_i}=y_{iT_i}\,|\,\mathbf{X}_i,\Sigma_{t=1}^{T_i}y_{it})=\frac{\exp\left[\Sigma_{t=1}^{T_i}y_{it}\boldsymbol{\gamma}'\mathbf{x}_{it}\right]}{\sum_{\Sigma_t d_{it}=\Sigma_t y_{it}}\exp\left[\Sigma_{t=1}^{T_i}y_{it}\boldsymbol{\gamma}'\mathbf{x}_{it}\right]}. \quad (2.51)$$

The function in the denominator is summed over the set of all $\begin{pmatrix} T_i \\ \Sigma_{t=1}^{T_i}y_{it}\end{pmatrix}$ different sequences of T_i zeros and ones that have the same sum as $\Sigma_{t=1}^{T_i}y_{it}$. (The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (2005, p. 235). In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

Consider the example of $T_i = 2$. The unconditional likelihood is

$L_i = \text{Prob}(Y_{i1} = y_{i1})\text{Prob}(Y_{i2} = y_{i2})$.

For each pair of observations, we have these possibilities:

1. $y_{i1} = 0$ and $y_{i2} = 0$. Prob(0,0 | sum = 0) = 1.
2. $y_{i1} = 1$ and $y_{i2} = 1$. Prob(1,1 | sum = 2) = 1.

The ith term in L_c for either of these is just one, so they contribute nothing to the conditional likelihood function. When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

$$3.\ \text{Prob}(0,1|\text{sum}=1)=\frac{\text{Prob}(0,1\text{ and sum}=1)}{\text{Prob(sum}=1)}=\frac{\text{Prob}(0,1)}{\text{Prob}(0,1)+\text{Prob}(1,0)}.$$

.

Therefore, for this pair of observations, the conditional probability is

$$\frac{\left(\frac{1}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i2})}\right)}{\left(\frac{1}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i2})}\right)+\left(\frac{\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i1})}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i1})}\frac{1}{1+\exp(\alpha_i+\boldsymbol{\gamma}'\mathbf{x}_{i2})}\right)}$$
$$=\frac{\exp(\boldsymbol{\gamma}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\gamma}'\mathbf{x}_{i1})+\exp(\boldsymbol{\gamma}'\mathbf{x}_{i2})}.$$

By conditioning on the sum of the two observations,we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0,1). Pairs of observations with (1,0) are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or T_i , constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods. [Cecchetti (1986) and Willis (2006) present an application of this model.]

Computation of partial effects in the fixed effects binary choice model presents a new problem. If the sample contains any groups that contain no variation – i.e., y_{it} is always one or zero – then those groups must be dropped from the sample. This is true both for the unconditional estimator or the Rasche/Chamberlain conditional estimator. (Note in the earlier results it is reported that 3,289 observations (groups) have been omitted from the sample.) This precludes computation of average partial effects for either estimator. This follows automatically in the conditional estimator since the constant terms are not computed. One might base estimation of partial effects on the individuals remaining in the sample. An alternative is to base the computation on the means of the data and the mean of the constant terms for the included groups. Then

$$\hat{\boldsymbol{\delta}} = f\left(\bar{\hat{\alpha}}_* + \hat{\boldsymbol{\gamma}}'\bar{\mathbf{x}}\right)\hat{\boldsymbol{\gamma}},$$

where $\bar{\mathbf{x}}$ would be the overall mean vector for the independent variables, or some suitably chosen alternative.

This does not solve the problem for the conditional estimator, however, since the constant terms are not estimated for that model. One way to proceed in this case is as follows: The log likelihood for one individual is

$$\ln L_i = \sum_{t=1}^{T_i} \ln \Lambda[q_i(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i)].$$

The problem solved by the conditional estimator is consistent estimation of γ. If γ were known, and if there were variation of y_i in the T_i observations, then estimation of α_i would be done by maximizing $\ln L_i$ with respect to α_i. Using the first order conditions derived earlier, we find the solution can be found by solving

$$p_{i1} = \frac{1}{T_i}\sum_{t=1}^{T_i} \Lambda\left(\alpha_i + a_{it}\right),$$

where p_{i1} is the proportion of the T_i observations with y_{it} equal to one and $a_{it} = \boldsymbol{\gamma}'\mathbf{x}_{it}$. There is no closed form solution, but the root can be found by a simple one dimensional search. Estimation of partial effects, probabilities, etc. can then be based on the average of these estimates.

Table 2.14 presents estimates of a fixed effects logit model, computed with both the conditional and unconditional estimators. The theory suggests that the coefficient estimates with the unconditional approach should be systematically larger than the conditional estimates. This does seem to be the case in general. The exception, the coefficient on Educ, is also the one with the lowest t ratio, by far. This suggests that for this coefficient, we should expect very large sampling variation. The major differences between the estimators show up in the partial effects. These are computed by obtaining estimates of the constant terms, then averaging over all observations. This is a large sample, so the sampling variabiity induced by the small T_i should be averaged away in the partial effects. We find that the effects computed with the two estimators are very different. With only the incidental parameters to provide guidance, we would opt for the estimates computed from the conditional estimator.

Table 2.14 Estimated Fixed Effects Logit Models

	Unconditional Estimator LogL = -8506.164 LogLR = -17365.15 7293 Individuals 3289 Individuals Bypassed				Conditional Estimator LogL = -5669.541				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
AGE	-.1095	.0076	-14.405	.0000	-.0881	.0068	-12.984	.0000	43.5257
EDUC	.0090	.0835	.108	.9141	.0126	.0718	.176	.8604	11.3206
INCOME	.6038	.1968	3.068	.0022	.4767	.1750	2.724	.0064	.35208
MARRIED	-.1091	.1114	-.979	.3276	-.0772	.0983	-.785	.4322	.75862
KIDS	-.0167	.0793	-.210	.8337	-.0059	.0706	-.084	.9331	.40273
	Partial Effects				Partial Effects				
AGE	-.0259	.0063	-4.102	.0000	-.0012	.00009	-13.961	.0000	43.5257
EDUC	.0021	.0193	.110	.9122	.0002	.0010	.176	.8605	11.3206
INCOME	.1429	.0582	2.455	.0141	.0066	.0023	2.920	.0035	.35208
MARRIED	-.0258	.0015	-17.531	.0000	-.0011	.0014	-.789	.4303	.75862
KIDS	-.0039	.0008	-5.225	.0000	-.00008	.0010	-.084	.9331	.40273

2.10.3 Random Effects

A specification that has the same structure as the random effects linear regression model has been implemented by Butler and Moffitt (1982) and is now in widespread use. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (2008a, Chapter 23). The random effects model specifies

$$\varepsilon_{it} = v_{it} + u_i,$$

where v_{it} and u_i are independent random variables with

$$E[v_{it}|\mathbf{X}] = 0;\ \mathrm{Cov}[v_{it}, v_{js}|\mathbf{X}] = \mathrm{Var}[v_{it}|\mathbf{X}] = 1, \text{ if } i = j \text{ and } t = s;\ 0 \text{ otherwise},$$
$$E[u_i|\mathbf{X}] = 0;\ \mathrm{Cov}[u_i, u_j|\mathbf{X}] = 0 \text{ if } i \neq j,\ \mathrm{Var}[u_i|\mathbf{X}] = \sigma_u^2,$$
$$\mathrm{Cov}[v_{it}, u_j|\mathbf{X}] = 0 \text{ for all } i, t, j,$$

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t. Then,

$$\mathrm{Cov}[\varepsilon_{it},\varepsilon_{is}] = \sigma_u^2$$

and

$$\mathrm{Corr}[\varepsilon_{it},\varepsilon_{is}] = \rho = \frac{\sigma_u^2}{1+\sigma_u^2}.$$

The new free parameter is

$$\sigma_u^2 = \rho/(1-\rho).$$

The Pooled Estimator

The implied probit model, given the composition of the disturbance is

$$y_{it}^* = \boldsymbol{\gamma}'\mathbf{x}_{it} + v_{it} + u_i,$$
$$y_{it} = 1(y_{it}^* > 0).$$

It follows that

$$\begin{aligned}\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it}) &= \text{Prob}(\boldsymbol{\gamma}'\mathbf{x}_{it} + v_{it} + u_i > 0) \\ &= \Phi\left(\frac{\boldsymbol{\gamma}'\mathbf{x}_{it}}{1+\sigma_u^2}\right) \\ &= \Phi\left(\boldsymbol{\gamma}_*'\mathbf{x}_{it}\right).\end{aligned} \tag{2.52}$$

If one pools the data and ignores the within group correlation, then the maximum likkelihood estimator provides a consistent estimator of γ_*, not $\boldsymbol{\gamma}$. So, the estimator is inconsistent; it is biased toward zero – *as an estimator of* $\boldsymbol{\gamma}$. Since the observations are correlated (within the groups), the estimated asymptotic covariance matrix will also be inappropriate. One would expect the cluster corrected covariance matrix estimator (see Section 2.10.1) to be an improvement. The partial effects in the random effects probit model, once again based on the preceding formulation, are precisely

$$\begin{aligned}\frac{\partial \text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it})}{\partial \mathbf{x}_{it}} &= \left[\phi(\boldsymbol{\gamma}_*'\mathbf{x}_{it})\right]\boldsymbol{\gamma}_* \\ &= \left[\phi\left(\sqrt{1-\rho}\right)\right)\left(\boldsymbol{\gamma}'\mathbf{x}_{it}\right)\Big]\Big[\left(\sqrt{1-\rho}\right)\right)\boldsymbol{\gamma}\Big].\end{aligned} \tag{2.53}$$

The implication is that although the pooled estimator does not estimate γ consistently, assuming the data, $\mathbf{x}_{it}$ are well behaved, *the pooled model does produce the appropriate estimator of the partial effects in the random effects probit model*. [Wooldridge (2002a) discusses this issue at some length.] The upshot would be that this establishes a case for estimating the pooled model, with an appropriate correction to the estimator of the asymptotic covariance matrix.

The Maximum Likelihood Estimator

The log likelihood for the random effects model is

$$\ln L = \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} F\left[q_{it}(\boldsymbol{\gamma}'\mathbf{x}_{it} + \sigma_u w_i)\right] f(w_i)\,dw_i, \tag{2.54}$$

where once again, $\mathbf{x}_{it}$ contains a constant term, and $w_i = u_i/\sigma_u$. Maximization of the log likelihood requires computation of the inner integrals, for which there is no closed form. Butler and Moffit's method based on Gauss-Hermite quadrature is a very common approach. The parameters may

also be estimated by maximum simulated likelihood. Details on both methods may be found in Greene (2008a, Chapter 23). This model is typically specified using the normal distribution (probit model) for both v_{it} and u_i. Using the simulation based estimator, the logit model could be used for either or both terms, though it is difficult to see a clear motivation for doing so.

As shown in (2.53), the partial effects in the model involve the scaling parameter (1-ρ). Since the MLE estimates γ and ρ, it follows that the MLE estimates the structural parameters consistently, but not the partial effects. In order to estimate partial effects based on the MLE, it is necessary to compute (2.53) using the estimators of γ and ρ.

GMM Estimation

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. For the more general panel probit model examined in Section 2.14, the GMM approach offers some savings in computational effort by avoiding evaluation of multivariate normal probabilities. For the random effects model considered here, the benefit is more limited, since the estimation requires only univariate normal integration.

Heckman and Singer's Semiparametric Approach

Heckman and Singer (1984a,b) argued that a fully parametric specification of the distribution of unobserved heterogeneity (in a duration model) could overspecify the model, and bias the estimation of the other parameters. Their proposed alternative is based on a discrete approximation to the underlying distribution of the individual heterogeneity. The Heckman and Singer model can be formulated as a latent class model. The implied latent class binary choice model is

$$\begin{aligned}&\text{Prob}(y_{it} = 1 \,|\mathbf{x}_{it}, \text{class} = c) \;=\; F(\alpha_c + \boldsymbol{\gamma}'\mathbf{x}_{it}),\\ &\text{Prob}(\text{class} = c) \;=\; \pi_c.\end{aligned} \tag{2.55}$$

(Note that we have isolated the constant term from the rest of the parameter vector.) The class probabilities are specified nonparametrically. The requirement that they be positive and sum to one can be imposed by a *multinomial logit* functional form

$$\pi_c \;=\; \frac{\exp(\theta_c)}{\Sigma_{c=1}^{C}\exp(\theta_c)},\; c = 1,...,C,\; \theta_C = 0.$$

Note that this function does not impose any restrictions on the probabilities other than that they are positive and sum to one. The C-1 parameters θ_c are unrestricted. The log likelihood function for this model is

$$\ln L = \sum_{i=1}^{n} \;\ln \sum_{c=1}^{C} \pi_c \prod_{t=1}^{T_i} \; F\left[q_{it}(\alpha_c + \boldsymbol{\gamma}'\mathbf{x}_{it})\right]. \tag{2.56}$$

The log likelihood function is maximized with respect to the C-1 class probabilities, C constant terms and K parameters in $\boldsymbol{\gamma}$.

Probabilities and partial effects for this model can be estimated in two ways. An unconditional approach can be based directly on the MLEs of the model parameters. Thus,

$$\begin{aligned}\text{Est.Prob}(y_{it}=1\,|\,\mathbf{x}_{it}) &= \sum_{c=1}^{C}\hat{\pi}_c F\left[\hat{\alpha}_c+\hat{\boldsymbol{\gamma}}'\mathbf{x}_{it}\right],\\ \hat{\boldsymbol{\delta}}_{it} &= \left\{\sum_{c=1}^{C}\hat{\pi}_c f\left[\hat{\alpha}_c+\hat{\boldsymbol{\gamma}}'\mathbf{x}_{it}\right]\right\}\boldsymbol{\gamma}.\end{aligned} \tag{2.57}$$

Alternatively, we can base an estimate of the class, c_i, within which the individual resides as follows, using Bayes theorem:

$$\begin{aligned}\text{Prob}(class=c\,|\,\mathbf{y}_i,\mathbf{X}_i) &= \frac{\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i},class=c\,|\,\mathbf{X}_i)}{\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i}\,|\,X_i)}\\ &= \frac{\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i}\,|\,\mathbf{X}_i,class=c)\text{Prob}(class=c)}{\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i}\,|\,\mathbf{X}_i)}\\ &= \frac{\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i}\,|\,\mathbf{X}_i,class=c)\text{Prob}(class=c)}{\sum_{c=1}^{C}\text{Prob}(\mathbf{y}_i=y_{i1},y_{i2}...y_{i,T_i}\,|\,\mathbf{X}_i,class=c)\text{Prob}(class=c)}\\ &= \frac{\pi_c\prod_{t=1}^{T_i}F\left[q_{it}(\alpha_c+\boldsymbol{\gamma}'\mathbf{x}_{it})\right]}{\sum_{c=1}^{C}\pi_c\prod_{t=1}^{T_i}F\left[q_{it}(\alpha_c+\boldsymbol{\gamma}'\mathbf{x}_{it})\right]}\\ &= \hat{\pi}_c\,|\,\mathbf{y}_i,\mathbf{X}_i.\end{aligned} \tag{2.58}$$

The natural estimator of which is the appropriate class for individual i would be the class with the largest conditional probability. Given this estimator of c_i, the estimator of α_c follows, then the probabilities and partial effects for individual *i* can be computed.

Table 2.15 presents estimates of a random effects model for *Healthy*. The left panel shows the Butler and Moffitt (1982) results using Gauss-Hermite quadrature for the integration. The panel on the right shows the same model estimated by maximum simulated likelihood. We used only 50 Halton draws for the simulation – one would typically use several hundred. Nonetheless, the estimates are surprisingly close. The implied estimate of ρ in the simulation is $\hat{\rho}=\hat{\sigma}_u^2/(1+\hat{\sigma}_u^2)=0.5412$, which differs only trivially from the quadrature based estimate.

Table 2.16 shows the estimates of Heckman and Singer's (1984a,b) semiparametric, latent class model. We begin with a specification search. There is no firm rule for determining the optimal number of classes. The likelihood ratio is not valid because a model with fewer classes is not parametrically nested in a larger one. In this specification, each class does require two additional parameters beyond the one lower. However, for example, one cannot produce a four class model by restricting the parameters of one of the classes. In principle, a four class model is produced from a five class model by forcing one of the αs to equal one of the other ones, but which one? And, if so, is there any restriction needed on the corresponding probability? As can be seen in the table, the log likelihood function does increase with the number of classes. However, there is no clear way to use this to formulate a search for the right number of classes. A common approach is to base the search on the lnAIC, which is lnAIC = (-2ln*L* + 2*M*)/*n*, where *M* is the number of model parameters. By this rule, it appears that the five class model is preferred. However, Heckman and Singer provide an additional suggestion which is useful here. They argue that if the model is fit with too many classes, the estimates will become unstable because the estimator (here) in an *M* dimensional parameter space is actually on a ridge in an *M*-2 (smaller) space. Note in Table 2.16, the estimate of α_1 is 10.238 with an estimated standard error of 66,988 and that for α_2 is -8.523 with standard error of 134,831. These would seem to fit their description. Thus, we chose the four class model as our preferred specification. The implied

mean and standard deviation for the discrete random variable in the four class model are 0.5933 and 1.1197. The standard deviation differs from the parametric estimate of 1.0862 by only about 3.1%. Though they are not directly comparable, it is striking that the log likelihood function for the four class latent class model is nearly identical to that for the parametric random effects model in Table 2.15.

Table 2.15 Estimated Random Effects Probit Models

	Quadrature Estimator				Simulation Estimator				
	LogL = -15424.40				LogL = -15429.26				
	LogL0 = -17365.76				LogL0 = -17365.76				
	7293 Individuals				Simulation = 50 Halton				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.9459	.1116	8.473	.0000	.9420	.0694	13.574	.0000	43.5257
AGE	-.0365	.0015	-24.279	.0000	-.0364	.0010	-37.801	.0000	43.5257
EDUC	.0817	.0073	11.230	.0000	.0815	.0044	18.742	.0000	11.3206
INCOME	.3207	.0717	4.474	.0000	.3225	.0547	5.899	.0000	.35208
MARRIED	.0188	.0346	.544	.5863	.0170	.0237	.716	.4741	.75862
KIDS	.0430	.0298	1.443	.1490	.0442	.0216	2.049	.0405	.40273
Rho	.5404	.0100	53.842	.0000					
Sigma_u					1.0862	.0129	84.119	.0000	

Table 2.16a Semiparametric Random Effects Probit Model*

5	4	3	2	1
α_1 10.238 (66988)	2.3363 (.1573)	2.2500 (.1400)	1.6015 (.1014)	.4963 (.0589)
α_2 -8.523 (134831)	-1.7635 (1.0297)	-.3831 (.1273)	-.0309 (.0986)	
α_3 .8291 (.1554)	-.1478 (.1660)	.8882 (.1292)		
α_4 -.2636 (.1426)	1.0104 (.1436)			
α_5 1.8971 (.1777)				
π_1 .0756	.2700	.3133	.5513	1.0000
π_2 .0207	.0323	.2429	.4487	
π_3 .3758	.2772	.4437		
π_4 .2404	.4205			
π_5 .2876				
lnL -15420.17	-15423.77	-15431.74	-15552.89	-17365.76
AIC 1.12963	1.12975	1.13019	1.13891	1.27145

* Estimated standard errors for α_c in parentheses

Table 2.16b Estimated Parameters for 4 Class Latent Class Model

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
AGE	-.0367	.0015	-24.110	.0000	43.5257
EDUC	.0795	.0076	10.426	.0000	11.3206
INCOME	.3340	.0743	4.491	.0000	.35206
MARRIED	.0122	.0359	.340	.7341	.75862
KIDS	.0485	.0304	1.596	.1104	.40273

2.10.4 Mundlak's Correction for the Probit and Logit Models

The incidental parameters problem is a compelling reason to be skeptical of the fixed effects estimator when T_i is small, as it is in our application. However, the assumption that the common effect u_i is uncorrelated with $\mathbf{x}_{it}$ is a disadvantage of the random effects model. An approach suggested by Mundlak (1978) and extended by Wooldridge (2002b) proposes a middle

ground between the two. We propose that the effects in the fixed effects model are projected on the means of the (time varying) regressors,

$$\alpha_i = \theta'\bar{\mathbf{x}}_i + w_i . \tag{2.59}$$

where w_i is normally distributed with mean zero and standard deviation σ_w and is uncorrelated with $\bar{\mathbf{x}}_i$ or with $\mathbf{x}_{it}$. (Wooldridge proposes that α_i be projected on all T vectors, $\mathbf{x}_{it}$ rather than on just the means. The practical problem with this approach shows up in our application, in which there is an unbalanced panel. Simply filling in the missing years with zeros is not a satisfactory solution; zero is not an appropraite value for the regressor vector in a given year.) Inserting (2.59) into the fixed effects formulation in Section 2.10.2 produces the modified model

$$y_{it}^* = \gamma'\mathbf{x}_{it} + \theta'\bar{\mathbf{x}}_i + \varepsilon_{it} + w_i, \tag{2.60}$$

which is a random effects model.

2.10.5 Testing for Heterogeneity

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i = \alpha$), there is no unusual problem, and the model can be estimated, as usual, as a pooled probit or logit model. The test is simple for the random effects model. A simple Wald (t) test of the statistical significance of the estimate of ρ is appropriate. Alternatively, one can use a likelihood ratio test by comparing the log likelihoods of the random effects and pooled models. The estimate of ρ in the estimated random effects model in Table 2.15 is 0.5404 with an estimated standard error of 0.001037. The implied t ratio of 53.8 is large enough to reject the hypothesis of homogeneity. Alternatively, we can use the likelihood ratio test. For the estimated random effects models, we have

$$\begin{aligned} \ln L_{Pooled} &= -17365.76, \\ \ln L_{RE} &= -15424.40, \\ \ln L_{Heckman,Singer} &= -15423.77. \end{aligned}$$

For testing in the parametric framework, the likelihood ratio statistic, with one degree of freedom, would be 2[(-15424.40) – (-17365.76)] = 3,882.72. This is far larger than the critical value of 3.84, so once again, the hypothesis is rejected. To use the semiparametric approach instead, we need to recalculate the degrees of freedom. The number of additional parameters that are estimated to produce the improvement in the log likelihood is three for the additional constant terms plus three for the unrestricted probabilities – the fourth is constrained so that they sum to one. The statistic is 2[(-15423.77) – (-17365.76)] = 3,883.98 with 6 degrees of freedom. The 95% critical value is 12.59, so the hypothesis of homogeneity is once again rejected.

Testing for heterogeneity in the fixed effects case is more difficult. Consider first the conditional logit approach. It is not possible to test the hypothesis using the likelihood ratio test because the two likelihoods are not comparable. The conditional likelihood is based on a restricted data set that excludes individuals for which y_{it} is the same in every period. Moreover, none of the usual tests of restrictions can be used because the individual effects are never actually estimated.

Hausman's (1978) specification test is a natural one to use here. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is

inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent, whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

$$\chi^2 = \left(\hat{\gamma}_{CML} - \hat{\gamma}_{ML}\right)' \left[Asy.Var\left(\hat{\gamma}_{CML}\right) - Asy.Var\left(\hat{\gamma}_{ML}\right)\right]^{-1} \left(\hat{\gamma}_{CML} - \hat{\gamma}_{ML}\right). \qquad (2.61)$$

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero. It might be tempting to eliminate from the sample at the outset groups of observations for which y_{it} is always zero or T_i. If so, then the samples used for the pooled estimator and the conditional MLE will be the same. However, there is now a danger that the resulting subsample used for the pooled model is choice based – See Section 2.15 – so that the pooled estimator would no longer be consistent even under the null hypothesis of homogeneity.

One cannot use this approach with the unconditional FE estimator. The reason is that the unconditional MLE is inconsistent even when the fixed effects model is correctly specified, because of the incidental parameters (small T) problem. Therefore, it would seem that there is a loose end in the econometric methodology; there is no appropriate for fixed effects vs. no effects for the probit model in the received literature.

2.10.6 Testing for Fixed or Random Effects: A Variable Addition Test

The usual approach of using the Hausman test to test for fixed vs. random effects in the linear model is unavailable here. The fixed effects maximum likelihood estimator is inconsistent under both the null and alternative hypotheses. The Wu (1973) *variable addition test* should be a viable alternative. In the Mundlak specification considered in the section 2.10.4, if the random effects model is appropriate, then the coefficients on the group means should be zero. If θ is not zero, this casts doubt on the random effects model, which suggests the fixed effects model as a preferable alternative. Like all such specification tests, the power of this procedure is uncertain; the simple FE model is not the only alternative that could produce a significant result. But, for these specific null and alternative hypotheses, the Wald and likelihood ratio tests should be usable. Table 2.16 presents the random effects and Mundlak estimates for the probit model. The coefficient estimates in the latter are noticeably different in the two models. The Wald statistic of 45.27922 and the likelihood ratio statistic of 40.280 are both far larger than the critical chi squared with 5 degrees of freedom, 11.07. This suggests that for these data, the fixed effects model is the preferred framework.

2.11 Parameter Heterogeneity

The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. We also examined a semiparametric approach based on a discrete distribution using Heckman and Singer's (1984a,b) approach. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model

is essentially semiparametric. It requires no specific distributional assumption, however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. Heckman and Singer's (1984b) model provides a less stringent model specification based on a discrete distribution of the latent heterogeneity.

Table 2.17 Random Effects Model with Mundlak Correction

	Random Effects Probit				Group Means Addition				
	LogL = -15424.40				LogL = -15404.26				
	LogL0 = -17365.76				LogL0 = -17365.76				
	7293 Individuals								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.9459	.1116	8.473	.0000	.6551	.1232	5.320	.0000	43.5257
AGE	-.0365	.0015	-24.279	.0000	-.0521	.0036	-14.582	.0000	43.5257
EDUC	.0817	.0073	11.230	.0000	.0031	.0421	.073	.9415	11.3206
INCOME	.3207	.0717	4.474	.0000	.2937	.0959	3.064	.0022	.35208
MARRIED	.0188	.0346	.544	.5863	-.0429	.0534	-.803	.4220	.75862
KIDS	.0430	.0298	1.443	.1490	-.0019	.0397	-.048	.9614	.40273
AGEBAR					.0193	.0039	4.895	.0000	
EDUCBAR					.0790	.0427	1.848	.0646	
INCMBAR					.3451	.1496	2.307	.0211	
MARRBAR					.0499	.0717	.695	.4871	
KIDSBAR					.0936	.0616	1.520	.1285	
Rho	.5404	.0100	53.842	.0000	.5389	.0100	53.822	.0000	

The preceding opens another possibility. The random effects model can be cast as a model with a random constant term;

$$y_{it}^* = \alpha_i + \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it}\,,\ i = 1, \ldots, n,\ t = 1, \ldots, Ti\,,$$
$$y_{it} = 1 \text{ if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where $\alpha_i = \alpha + \sigma w_i$. This is simply a reinterpretation of the model just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y_{it}^* = \alpha_i + \boldsymbol{\gamma}_i'\mathbf{x}_{it} + \varepsilon_{it}\,,\ i = 1, \ldots, n,\ t = 1, \ldots, Ti\,, \tag{2.62}$$
$$y_{it} = 1 \text{ if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where $\boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \boldsymbol{\Sigma}\mathbf{w}_i$ where $\boldsymbol{\Sigma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of maximum simulated likelihood is well suited to this model. The simulated log-likelihood for the random parameters model is

$$\ln L_S = \sum_{i=1}^{n} \ln \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} F\{q_{it}[(\boldsymbol{\gamma} + \boldsymbol{\Sigma}\mathbf{w}_{i,r})'\mathbf{x}_{it}]\}\,. \tag{2.63}$$

The simulation now involves R draws from the multivariate distribution of $\mathbf{u}$. Because the draws are uncorrelated - $\boldsymbol{\Sigma}$ is diagonal - this is essentially the same estimation problem as the random effects model considered previously. The simulated log likelihood is maximized with respect to the elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$. An interesting, relatively straightforward extension is to relax the assumption that the random parameters are uncorrelated. This can be done by writing $\boldsymbol{\Sigma} = \mathbf{LL}'$ where $\mathbf{L}$ is a lower triangular matrix, then simply including the below diagonal elements of $\mathbf{L}$ among the parameters to be estimated. A hierarchical model is obtained by allowing the parameter heterogeneity to be partly systematic, in terms of observed variables, as in

$$\boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Sigma}\mathbf{w}_i.$$

where $\boldsymbol{\Delta}$ is a matrix of parameters and $\mathbf{z}_i$ is a vector of covariates. The techniques are illustrated in the following example, in which the hierarchical model is specified to allow both random heterogeneity in the parameters and variation across genders.

An extension of the heterogeneity model to the latent class structure is a minor extension of the Heckman and Singer model of Section 2.10.3. We can also produce a counterpart to the hierarchical model as shown in (2.64) and (2.65). The model structure is

$$\pi_{ic} = \frac{\exp(\boldsymbol{\theta}_c'\mathbf{z}_i)}{\Sigma_{c=1}^{C}\exp(\boldsymbol{\theta}_c'\mathbf{z}_i)},\ c = 1,...,C,\ \boldsymbol{\theta}_C = \mathbf{0}, \tag{2.64}$$

$$\ln L = \sum_{i=1}^{n} \ln \sum_{c=1}^{C} \pi_{ic} \prod_{t=1}^{T_i} F\left[q_{it}(\boldsymbol{\gamma}_c'\mathbf{x}_{it})\right]. \tag{2.65}$$

Estimation of a fully specified latent class model is discussed in Section 8.2.5, and Greene (2008a). Background material on latent class models may be found in Mcachlan and Peel (2000) and Greene (2008, Section 16.9.7).

We have reestimated the probit model for *Healthy* with the basic random parameters (RP) specification,

$$\gamma_{ik} = \gamma_k + \sigma_k w_{ik}.$$

The results are shown in Table 2.18 with the original (now fixed parameters) estimates. The estimated means of the RPs differ substantially from their fixed counterparts. The differences can be seen in the first column of the table of partial effects in Table 2.19 as well. A likelihood ratio test of the null hypothesis of the fixed parameters model gives a chi squared value of 2(-15407.51 – (-17365.76)) = 3916.50 with six degrees of freedom. The hypothesis would be rejected, suggesting that the RP model is the preferred specification.

Partial effects in the RP model will vary both with the data and by the individual;

$$\frac{\partial \text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it})}{\partial \mathbf{x}_{it}} = f(\boldsymbol{\gamma}_i'\mathbf{x}_{it})\boldsymbol{\gamma}_i, \text{ where } \boldsymbol{\gamma}_i = \boldsymbol{\gamma} + \mathbf{L}\mathbf{w}_i.$$

This cannot be computed because it involves the unknown random term, $\mathbf{w}_i$. The simplest way to remove the indeterminacy in the computation is to use the population mean, $E[\boldsymbol{\gamma}_i] = \boldsymbol{\gamma}$ in the computation. An alternative approach is to estimate $\boldsymbol{\gamma}_i$ for the individual – in principle this would also allow the computation of average partial effects. Since $\mathbf{w}_i$ is random and uncorrelated with the observed variables, it is not possible to estimate $\boldsymbol{\gamma}_i$, itself. It is possible to improve on $E[\boldsymbol{\gamma}_i]$, hwever. Let $\mathbf{y}_i = (y_{i1},\ldots,y_{i,Ti})$ and let $\mathbf{X}_i$ denote the T_i observations on $\mathbf{x}_{it}$. Then, using Bayes Theorem,

$$f(\boldsymbol{\gamma}_i \mid \mathbf{y}_i, \mathbf{X}_i) = \frac{f(\boldsymbol{\gamma}_i, \mathbf{y}_i \mid \mathbf{X}_i)}{f(\mathbf{y}_i \mid \mathbf{X}_i)}$$
$$= \frac{f(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{X}_i) p(\boldsymbol{\gamma}_i)}{f(\mathbf{y}_i \mid \mathbf{X}_i)}$$
$$= \frac{f(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{X}_i) p(\boldsymbol{\gamma}_i)}{\int_{\boldsymbol{\gamma}_i} f(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{X}_i) p(\boldsymbol{\gamma}_i) d\boldsymbol{\gamma}_i}.$$

This is the conditional density of $\boldsymbol{\gamma}_i$ given the data on individual i. This provides a method of estimating the expectation of $\boldsymbol{\gamma}_i$ or of $\partial \text{Prob}(y_{it} = 1|\mathbf{x}_{it})/\partial \mathbf{x}_{it}$,

$$E(\boldsymbol{\gamma}_i \mid \mathbf{y}_i, \mathbf{X}_i) = \frac{\int_{\boldsymbol{\gamma}_i} \boldsymbol{\gamma}_i f(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{X}_i) p(\boldsymbol{\gamma}_i) d\boldsymbol{\gamma}_i}{\int_{\boldsymbol{\gamma}_i} f(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{X}_i) p(\boldsymbol{\gamma}_i) d\boldsymbol{\gamma}_i}. \tag{2.66}$$

The conditional mean of the partial effect would be obtained likewise. The integrals needed to obtain the result will not exist in closed form, but they can be simulated. In (2.66), $f(\mathbf{y}_i|\gamma_i,\mathbf{X}_i)$ is the contribution of individual i to the likelihood function (not its log) and $p(\boldsymbol{\gamma}_i)$ is the marginal density of $\boldsymbol{\gamma}_i$ which we have assumed is normal with mean $\boldsymbol{\gamma}$ and variance $\boldsymbol{\Sigma}$. Using the definition of the likelihood function in (2.9), then, the empirical counterpart to (2.66) is

$$E(\boldsymbol{\gamma}_i \mid \mathbf{y}_i, \mathbf{X}_i) = \frac{\int_{\boldsymbol{\gamma}_i} \boldsymbol{\gamma}_i \prod_{t=1}^{T_i} F(q_{it} \boldsymbol{\gamma}_i' \mathbf{x}_{it}) N_K(\boldsymbol{\gamma}_i \mid \boldsymbol{\gamma}, \boldsymbol{\Sigma}) d\boldsymbol{\gamma}_i}{\int_{\boldsymbol{\gamma}_i} \prod_{t=1}^{T_i} F(q_{it} \boldsymbol{\gamma}_i' \mathbf{x}_{it}) N_K(\boldsymbol{\gamma}_i \mid \boldsymbol{\gamma}, \boldsymbol{\Sigma}) d\boldsymbol{\gamma}_i}.$$

Since even given the population values of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$, the integrals would not be directly computable, we will use simulation instead. Inserting our estimates of the population population parameters, then, the estimator is

$$\hat{E}(\boldsymbol{\gamma}_i \mid \mathbf{y}_i, \mathbf{X}_i) = \frac{\frac{1}{R} \sum_{r=1}^{R} (\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r}) \prod_{t=1}^{T_i} F[q_{it} (\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r})' \mathbf{x}_{it}]}{\frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} F[q_{it} (\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r})' \mathbf{x}_{it}]} \tag{2.67}$$

where $\mathbf{LL}' = \boldsymbol{\Sigma}$ and $\mathbf{w}_{i,r}$ is the rth simulated draw from the K-variate standard normal population. If the random parameters are uncorrelated, then $\mathbf{L}$ is the diagonal matrix of estimated standard deviations. This provides an individual specific estimate, though we emphasize, it is the conditional mean function, not a direct estimate of $\boldsymbol{\gamma}_i$. It is the minimum mean squared error predictor of $\boldsymbol{\gamma}_i$ given $\mathbf{y}_i$ and $\mathbf{X}_i$. We have done this computation for our estimated random parameters model in Table 2.18. A kernel density estimator for γ_{INCOME} based on the 7,293 estimates is shown in Figure 2.8. In order to simulate the partial effects, the initial term $(\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r})$ in the numerator of (2.67) is replaced with $f[(\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r})' \mathbf{x}_{it}](\hat{\boldsymbol{\gamma}} + \hat{\mathbf{L}} \mathbf{w}_{i,r})$. These are the values shown in Table 2.19 for the two random parameters models.

The estimator in (2.67) is the counterpart to a Bayesian posterior mean. [Suggestions that the Bayesian estimator provides individual parameter estimates, $\boldsymbol{\gamma}_i$, as in Rossi and Allenby (1999) are in error; the Bayesian estimator also estimates the conditional mean function,

$E[\boldsymbol{\gamma}_i|\mathbf{Data}_i]$, not $\boldsymbol{\gamma}_i$, itself.] One difference between the classical estimator in (2.67) and the Bayesian estimator is that (2.67) treats the estimated structural parameters as if they were known. One could use (2.67) to estimate $\text{Var}[\gamma_{i,k}|\mathbf{y}_i,\mathbf{X}_i]$ by simulating an estimator of the expected square, then constructing a variance or standard deviation. This could form the basis of a 'confidence interval' for $\gamma_{i,k}$, which would be somewhat too narrow because it would ignore the sampling variability in the estimators of the structural parameters, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$. One possibility to reconcile this would be to bootstrap the interval over the estimated asymptotic distribution of the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$. The narrower Bayesian HPD interval would follow from the fact that the Bayesian estimator is posterior only to the data in hand while the classical estimator with its asymptotic variance attempts to characterize the entire population.

Table 2.18 Estimated Random Parameter Models

	Random Parameters				Pooled				
	LogL = -15407.51				LogL = -17365.76				
	LogLR = -17365.76				LogL0 = -18279.95				
	7293 Individuals								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
	Mean Parameters in Probability								
Constant	.5688	.0701	8.113	.0000	.4963	.0589	8.425	.0000	1.0000
AGE	-.0355	.0010	-37.027	.0000	-.0232	.0008	-28.991	.0000	43.5257
EDUC	.1103	.0046	23.839	.0000	.0573	.0037	15.467	.0000	11.3206
INCOME	.3137	.0554	5.665	.0000	.3425	.0481	7.118	.0000	.35208
MARRIED	.0265	.0237	1.117	.2641	.0129	.0206	.627	.5307	.75862
KIDS	.0560	.0219	2.563	.0104	.0666	.0186	3.581	.0003	.40273
	Variance Parameters in Random Parameter Distribution								
Constant	.0474	.0091	5.223	.0000					1.0000
AGE	.0144	.0002	60.817	.0000					43.5257
EDUC	.0778	.0011	71.857	.0000					11.3206
INCOME	.1934	.0240	8.068	.0000					.35208
MARRIED	.0205	.0106	1.939	.0525					.75862
KIDS	.3771	.0153	24.645	.0000					.40273

Table 2.19 Estimated Partial Effects

Variable	Est.	S.E.	t	P	Est.	S.E.	t	P	Mean of X
	Pooled				**Random Effects**				
AGE	-.0089	.0003	-29.012	.0000	-.0133	.0003	-43.726	.0000	43.5257
EDUC	.0219	.0014	15.478	.0000	.0297	.0023	12.957	.0000	11.3206
INCOME	.1309	.0184	7.118	.0000	.1175	.0204	5.763	.0000	.35208
MARRIED	.0049	.0079	.626	.5311	.0062	.0087	.715	.4747	.75862
KIDS	.0254	.0071	3.589	.0003	.0161	.0080	2.007	.0448	.40273
	Uncorrelated Random Parameters				**Correlated Random Parameters**				
AGE	-.0130	.0004	-34.711	.0000	-.0134	.0004	-34.044	.0000	43.5257
EDUC	.0402	.0017	23.757	.0000	.0390	.0018	21.443	.0000	11.3206
INCOME	.1144	.0202	5.670	.0000	.1158	.0218	5.303	.0000	.35208
MARRIED	.0097	.0087	1.115	.2648	.0096	.0092	1.043	.2971	.75862
KIDS	.0204	.0079	2.588	.0097	.0121	.0083	1.457	.1451	.40273

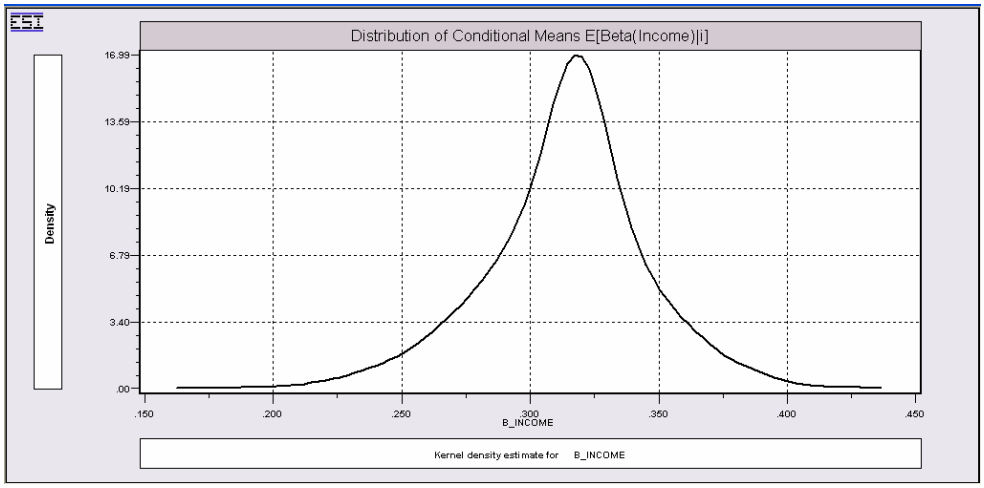


Figure 2.8 Distribution of Conditional Means of Income Parameter

2.12 Endogeneity of a RHS variable

The presence of endogenous right-hand-side variables in a binary choice model presents familiar problems for estimation. The problem is made worse in nonlinear models because even if one has an instrumental variable readily at hand, it may not be immediately clear what is to be done with it. The usual instrumental variable estimator is based on moments of the data, variances and covariances. In this binary choice setting, we are not using any form of least squares to estimate the parameters, so the IV method would appear not to apply. Consider the model

$$y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i + \varepsilon_i,$$
$$y_i = 1(y_i^* > 0),$$
$$E[\varepsilon_i \mid h_i] = g(h_i) \neq 0.$$

Thus, h_i is endogenous in this model. The simple maximum likelihood estimators considered earlier will not consistently estimate $(\boldsymbol{\gamma},\theta)$. [Without an additional specification that allows us to formalize Prob$(y_i = 1 \mid \mathbf{x_i}, h_i)$, we cannot state what the MLE will, in fact, estimate.] Suppose that we have a "relevant" instrumental variable, z_i such that

$$E[\varepsilon_i \mid z_i, \mathbf{x}_i] = 0,$$
$$E[h_i z_i] \neq 0.$$

A natural instrumental variable estimator would be based on the "moment" condition

$$E\left[\left(y_i^* - \boldsymbol{\gamma}'\mathbf{x}_i - \theta h_i\right)\begin{pmatrix} x_i \\ z_i \end{pmatrix}\right] = \mathbf{0}.$$

However, y_i^* is not observed, y_i is. The "residual," $\left(y_i^* - \boldsymbol{\gamma}'\mathbf{x}_i - \theta h_i\right)$, would have no meaning even if the true parameters were known. One approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variable is orthogonal to the residual $[\left(y_i - \Phi(\boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i)\right).]$; that is,

$$E\left[\left(y_i - \Phi(\boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i)\right)\begin{pmatrix}\mathbf{x}_i \\ z_i\end{pmatrix}\right] = \mathbf{0}.$$

This form of the moment equation, based on observables, can form the basis of a straightforward two-step GMM estimator.

The GMM estimator is not less parametric than the full information maximum likelihood estimator described below because the probit model based on the normal distribution is still invoked to specify the moment equation. Nothing is gained in simplicity or robustness of this approach to full information maximum likelihood estimation, which we now consider. (As Bertschek and Lechner argue, however, the gains might come in terms of practical implementation and computation time. The same considerations motivated Avery et al.)

The maximum likelihood estimator requires a full specification of the model, including the assumption that underlies the endogeneity of h_i. The model equations are

$$\begin{aligned}
&y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i + \varepsilon,\ y_i = 1(y_i^* > 0),\\
&h_i = \boldsymbol{\alpha}'\mathbf{c}_i + u_i,\\
&\begin{pmatrix}\varepsilon_i \\ u_i\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2\end{pmatrix}\right].
\end{aligned}$$

(We are assuming that there is a vector of instrumental variables, $\mathbf{c}_i$.) Probit estimation based on y_i and ($\mathbf{x}_i$,h_i) will not consistently estimate (γ, θ) because of the correlation between h_i and ε_i induced by the correlation between u_i and ε_i. Several methods have been proposed for estimation of this model. One possibility is to use the partial reduced form obtained by inserting the second equation in the first. This becomes a probit model with probability

$$\text{Prob}(y_i=1|\mathbf{x}_i,\mathbf{c}_i) = \Phi(\boldsymbol{\gamma}^{*\prime}\mathbf{x}_i + \boldsymbol{\alpha}^{*\prime}\mathbf{c}_i).$$

This will produce consistent estimators of

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma} / \sqrt{1 + \theta^2\sigma_u^2 + 2\theta\rho\sigma_u}$$

and

$$\boldsymbol{\alpha}^* = \theta\boldsymbol{\alpha} / \sqrt{1 + \theta^2\sigma_u^2 + 2\theta\rho\sigma_u}$$

as the coefficients on $\mathbf{x}_i$ and $\mathbf{c}_i$, respectively. (The procedure will estimate a mixture of $\boldsymbol{\gamma}^*$ and $\boldsymbol{\alpha}^*$ for any variable that appears in both $\mathbf{x}_i$ and $\mathbf{c}_i$.) In addition, linear regression of h_i on $\mathbf{c}_i$ produces estimates of $\boldsymbol{\alpha}$ and σ_u^2, but there is no method of moments estimator of ρ or θ produced by this procedure, so this estimator is incomplete. Newey (1987) suggested a "minimum chi-squared" estimator that does estimate all parameters. A more direct, and actually simpler approach is full information maximum likelihood.

The log-likelihood is built up from the joint density of y_i and h_i , which we write as the product of the conditional and the marginal densities,

$$f(y_i,h_i) = f(y_i|h_i)\,f(h_i).$$

To derive the conditional distribution, we use results for the bivariate normal, and write

$$\varepsilon_i|u_i = [(\rho\sigma_u)/\sigma_u^2]u_i + v_i,$$

where v_i is normally distributed independently of u_i with zero mean and Var[v_i] = (1 − ρ^2). Inserting this in the first equation, we have

$$y_i^*|h_i = \boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i + (\rho/\sigma_u)u_i + v_i.$$

Therefore,

$$\text{Prob}(y_i = 1|\mathbf{x}_i,h_i) = \Phi\left[\frac{\boldsymbol{\gamma}'\mathbf{x}_i + \theta h_i + (\rho/\sigma_u)u_i}{\sqrt{1-\rho^2}}\right]. \tag{2.68}$$

Inserting the expression for u_i = (h_i − $\boldsymbol{\alpha}'\mathbf{c}_i$), and using the normal density for the marginal distribution of h_i in the second equation, we obtain the log-likelihood function for the sample,

$$\ln L = \sum_{i=1}^{n} \ln\Phi\left[q_i\left(\frac{\gamma' x_i + \theta h_i + (\rho/\sigma_u)(h_i - \boldsymbol{\alpha}'\mathbf{c}_i)}{\sqrt{1-\rho^2}}\right)\right] + \ln\left[\frac{1}{\sigma_u}\phi\left(\frac{h_i - \boldsymbol{\alpha}'\mathbf{c}_i}{\sigma_u}\right)\right]. \tag{2.69}$$

(A built-in Stata procedure to compute this maximum likelihood estimator is unfortunately labeled "IVPROBIT" giving users the impression that it is using an instrumental variables estimator rather than the full information MLE.)

The case in which the endogenous variable in the main equation is, itself, a binary variable occupies a large segment of the literature. Consider the model

$$\begin{aligned} &y_i^* = \boldsymbol{\gamma}'\mathbf{x}_i + \theta T_i + \varepsilon, \\ &y_i = 1(y_i^* > 0), \\ &E[\varepsilon_i \mid T_i] = 0, \end{aligned}$$

where T_i is a binary variable indicating some kind of program participation (e.g., graduating from high school or college, receiving some kind of job training, etc.). The model in this form (and several similar ones) is a *treatment effects model*. The main object of estimation is θ (at least superficially). In these settings, the observed outcome may be y_i^* (e.g., income or hours) or y_i (e.g., labor force participation). The preceding analysis has suggested that problems of endogeneity will intervene in either case.

2.13 Bivariate Probit Models

A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned} &y_1^* = \boldsymbol{\gamma}_1'\mathbf{x}_1 + \varepsilon_1, \quad y_1 = 1(y_1^* > 0), \\ &y_2^* = \boldsymbol{\gamma}_2'\mathbf{x}_2 + \varepsilon_2, \; y_2 = 1(y_2^* > 0), \\ &\text{E}[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = \text{E}[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 0, \\ &Var[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = Var[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 1, \\ &Cov[\varepsilon_1, \varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = \rho. \end{aligned} \tag{2.70}$$

There is no convenient formulation of the bivariate choice model based on the logistic distribution. The bivariate probit (normal) specification is used with only rare exception in applications. The bivariate normal cdf is

$$\text{Prob}(X_1 < x_1,\ X_2 < x_2) \ = \ \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho) dz_1 dz_2,$$

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is

$$\phi_2(x_1, x_2, \rho) = \frac{\exp[-\frac{1}{2}(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)]}{2\pi\sqrt{1-\rho^2}}. \tag{2.71}$$

To construct the log-likelihood, let $q_{i1} = 2y_{i1} - 1$ and $q_{i2} = 2y_{i2} - 1$. Thus, $q_{ij} = +1$ if $y_{ij} = 1$ and -1 if $y_{ij} = 0$ for $j = 1$ and 2. Now let $z_{ij} = \boldsymbol{\gamma}_j' \mathbf{x}_{ij}$, $w_{ij} = q_{ij}\ z_{ij}$, $j = 1,\ 2,$ and $\rho_i^* \ = q_{i1}q_{i2}\rho$. Note the notational convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates the variables in the second equation. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf. The probabilities that enter the likelihood function are

$$\text{Prob}(Y_{i1} = y_{i1},\ Y_{i2} = y_{i2} | \mathbf{x}_{i1}, \mathbf{x}_{i2}) \ = \ \Phi_2(w_{i1}.w_{i2}, \rho_i^*),$$

which accounts for all the necessary sign changes needed to compute probabilities for y's equal to zero and one. Thus,

$$\begin{aligned} \ln L \ &= \ \sum_{i=1}^{n} \ \ln \Phi_2(w_{i1}.w_{i2}, \rho_i^*) \\ &= \ \sum_{i=1}^{n} \ \ln \Phi_{i2}. \end{aligned} \tag{2.72}$$

The derivatives of the log-likelihood then reduce to

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\gamma}_j} &= \sum_{i=1}^{n} \ \left(\frac{q_{ij} g_{ij}}{\Phi_{i2}} \right) \mathbf{x}_{ij}, j = 1, 2, \\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=1}^{n} \ \frac{q_{i1} q_{i1} \phi_{i2}}{\Phi_{i2}}, \end{aligned} \tag{2.73}$$

where

$$g_{i1} = \phi(w_{i1}) \Phi \left[\frac{w_{i2} - \rho w_{i1}}{\sqrt{1-\rho^2}} \right], \tag{2.74}$$

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. It is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L / \partial \boldsymbol{\gamma}_1$, if $\rho = \rho_i^* = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in the partial derivatives with q_{i1} and q_{i2} produces the results for the univariate probit model. Because both functions in $\partial \ln L / \partial \rho$ factor into the product of the univariate functions, $\partial \ln L / \partial \rho$ reduces to

$$\frac{\partial \ln L}{\partial \rho} | (\rho = 0) = \sum_{i=1}^{n} \left(\frac{q_{i1}\phi(q_{i1}\boldsymbol{\gamma}_1'\mathbf{x}_{i1})}{\Phi(q_{i1}\boldsymbol{\gamma}_1'\mathbf{x}_{i1})} \right) \left(\frac{q_{i2}\phi(q_{i2}\boldsymbol{\gamma}_2'\mathbf{x}_{i2})}{\Phi(q_{i2}\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \right) \tag{2.75}$$
$$= \sum_{i=1}^{n} \lambda_{i1}\lambda_{i2}.$$

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. Computation of the bivariate normal integrals needed for the log likelihood function can be done using quadrature methods. Expressions for the second derivatives to use for computing an asymptotic covariance matix for the MLE are given in Greene (2008a, p. 819). Given the complexity of the expressions, this seems like an opportune point to use the Berndt, Hall, Hall and Hausman estimator based on only the first derivatives.

2.13.1 Tetrachoric Correlation

The tetrachoric correlation is the correlation coefficient computed for a pair of binary variables that are assumed to be derived by censoring two observations from an underlying continuous bivariate normal population. This would be the bivariate probit model without independent variables. In this representation, the *tetrachoric correlation* is precisely the ρ in this model - it is the correlation that would be measured between the underlying continuous variables if they could be observed. This suggests an interpretation of the correlation coefficient in a bivariate probit model—as the *conditional tetrachoric correlation*. It also suggests a method of easily estimating the tetrachoric correlation coefficient using a program that is built into nearly all commercial software packages. We obtain an estimate of ρ simply by fitting a bivariate probit model with no covariates.

In the example below, we will analyze the variable *Working* in a bivariate probit model with *Healthy*. (The analysis will be based on the 4,483 observations used in the previous examples.) A cross tabulation for these two variables appears in Table 2.18. The simple (Pearson) correlation between these two binary variables is 0.09288. The tetrachoric correlation computed from a bivaraite probit model is 0.15159.

Table 2.20 Cross Tabulation of Healthy and Working

Chi-squared[1] = 38.76382 (Prob = 0.00000)			
	WORKING		
HEALTHY	0	1	Total
0	678 (0.151)	1099 (0.245)	1777 (0.396)
1	791 (0.176)	1915 (0.428)	2706 (0.604)
Total	1469 (0.327)	3014 (0.763)	4483 (1.000)

2.13.2 Testing for Zero Correlation

The Wald and likelihood ratio tests are the usual devices for testing the hypothesis that ρ equals zero in the bivariate probit model. For the Wald test, the square of the *t* statistic for $\hat{\rho}$ presented with the standard output has a limiting chi squared distribution with one degree of freedom.. For the example in Table 2.21, $\hat{\rho}$ = .0572 with a reported *t* statistic of 2.08. The chi squared value is 4.3264, which leads us to reject the hypothesis for this specification. The likelihood ratio test is carried out y comparing the log likelihood function for the bivariate probit

model to the sum of the seaprate log likelihoods for the univariate probits that are implied when ρ = 0. The statistic is

$$LR = 2[\ln L_{Bivariate} - (\ln L_1 + \ln L_2)].$$

The statistic has one degree of freedom. For the example, the result is

$$LR = 2[-5294.053 - (-2890.288 + -2405.931)] = 4.332.$$

The hypothesis is rejected once again. Both of these statistics require estimation of the bivariate probit model. The Lagrange multiplier test derived by Kiefer (1982) is based on only the single equation results. The statistic is

$$LM = \frac{\left[\sum_{i=1}^{n} \frac{q_{i1}q_{i2}\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})}\right]^2}{\sum_{i=1}^{n} \frac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(w_{i2})\Phi(-w_{i1})\Phi(-w_{i2})}}$$

For the data and single equation estimates (not shown) for the model in Table in 2.21, the statistic equals 4.0956. As in the other two cases, we reject the hypothesis that ρ equals zero.

2.13.3 Marginal Effects in a Bivariate Probit Model

There are several possible marginal effects one might want to evaluate in a bivariate probit model. [See Greene (1996, 2008a) and Christofides et al. (1997, 2000).] A natural first step would be the derivatives of Prob[$y_1 = 1, y_2 = 1 \mid \mathbf{x}_1, \mathbf{x}_2$] in (2.73). These can be deduced from $\partial \ln L/\partial \boldsymbol{\gamma}_j$ by multiplying by Φ_2, removing the sign carrier, q_{ij} and differentiating with respect to $\mathbf{x}_j$ rather than $\boldsymbol{\gamma}_j$. The result is

$$\frac{\partial \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1, \boldsymbol{\gamma}_2'\mathbf{x}_2, \rho)}{\partial \mathbf{x}_1} = \phi(\boldsymbol{\gamma}_1'\mathbf{x}_1)\Phi\left(\frac{\boldsymbol{\gamma}_2'\mathbf{x}_2 - \rho\boldsymbol{\gamma}_1'\mathbf{x}_1}{\sqrt{1-\rho^2}}\right)\boldsymbol{\gamma}_1. \tag{2.76}$$

The bivariate probability, albeit possibly of interest in its own right, is not a conditional mean function. As such, the preceding does not correspond to a regression coefficient or a slope of a conditional expectation. For convenience in evaluating the conditional mean and its partial effects, we will define a vector $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}2$ and let $\boldsymbol{\gamma}_1'\mathbf{x}_1 = \boldsymbol{\beta}_1'\mathbf{x}$. Thus, $\boldsymbol{\beta}_1$ contains all the nonzero elements of $\boldsymbol{\gamma}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; β_2 is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}] = \Phi(\boldsymbol{\beta}_1'\mathbf{x}, \boldsymbol{\beta}_2'\mathbf{x}, \rho).$$

Signs are changed appropriately if the probability of the zero outcome is desired in either case. The marginal effects of changes in $\mathbf{x}$ on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\boldsymbol{\beta}_1 + g_2\boldsymbol{\beta}_2, \tag{2.77}$$

where g_1 and g_2 were defined in (2.74). The familiar univariate cases will arise if ρ =0, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector.

There are also some regression functions to consider. The unconditional mean functions are given by the univariate probabilities:

$$E[y_j|\mathbf{x}_1,\mathbf{x}_2] = \Phi(\boldsymbol{\gamma}_j'\mathbf{x}_j), j = 1,2,$$

which was analyzed in detail earlier. One pair of conditional mean functions that might be of interest are

$$\begin{aligned} E[y_1 \mid y_2 = 1, \mathbf{x}] &= \text{Prob}(y_1 = 1 \mid y_2 = 1, \mathbf{x}) \\ &= \frac{\text{Prob}(y_1 = 1, y_2 = 1 \mid \mathbf{x})}{\text{Prob}(y_2 = 1 \mid \mathbf{x})} \\ &= \frac{\Phi_2(\boldsymbol{\beta}_1'\mathbf{x}, \boldsymbol{\beta}_2'\mathbf{x}, \rho)}{\Phi(\boldsymbol{\beta}_2'\mathbf{x})}, \end{aligned} \tag{2.78}$$

and similarly for $E[y_2 \mid y_1 = 1, \mathbf{x}]$. The marginal effects for this function are given by

$$\frac{\partial E[y_1 \mid y_2 = 1, x]}{\partial \mathbf{x}} = \left(\frac{1}{\Phi(\boldsymbol{\beta}_2'\mathbf{x})}\right)\left[g_1\boldsymbol{\beta}_1 + \left(g_2 - \Phi_2\frac{\phi(\boldsymbol{\beta}_2'\mathbf{x})}{\Phi(\boldsymbol{\beta}_2'\mathbf{x})}\right)\boldsymbol{\beta}_1\right]. \tag{2.79}$$

Finally, one might construct the nonlinear conditional mean function

$$E[y_1 \mid y_2, \mathbf{x}] = \frac{\Phi_2(\boldsymbol{\beta}_1'\mathbf{x}, (2y_2 - 1)\boldsymbol{\beta}_2'\mathbf{x}, (2y_2 - 1)\rho)}{\Phi[(2y_2 - 1)\boldsymbol{\beta}_2'\mathbf{x}]}. \tag{2.80}$$

The derivatives of this function are the same as those presented earlier, with sign changes in several places if $y2 = 0$ is the argument.

In each of these sets of partial effects, there is a direct and an indirect effect of a changing variable. The direct effect is the effect on E[y_1…] of a variable that appears in $\mathbf{x}_1$. The indirect effect is the effect of a variable that appears in $\mathbf{x}_2$. When variables appear in both equations, the total effect will be the sum of the two effects. In the example below, for example, *Age* and *Educ* appear in both equations, so there is a decomposition of the partial effects for each of these.

We have added a second equation to the probit model for *Healthy*,

$$\text{Prob}(Working_i = 1|\mathrm{x}_{i,Working}) = F(Age, Educ, Female).$$

Estimates of the bivariate probit model are shown in Table 2.21. The conditional tetrachoric correlation between these two variables is statistically significant, but quite small. (Three tests of the significance are carried out in Section 2.13.2.) The partial effects for Prob(Healthy=1|Working=1) are shown at the left of the table. The partial effects can be decomposed into direct and indirect effects for variables that appear in both equations, *Age* and *Educ*.

Table 2.21 Estimated Bivariate Probit Model

	HEALTHY [lnL=-2890.288] lnL = -5294.053				WORKING [lnL = -2405.931] (n = 4483)				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.4814	.1419	3.392	.0007	1.3418	.1550	8.658	.0000	1.0000
AGE	-.0203	.0020	-10.335	.0000	-.0223	.0017	-12.867	.0000	43.4401
EDUC	.0531	.0087	6.099	.0000	.0551	.0099	5.574	.0000	11.4181
INCOME	.1602	.1280	1.251	.2109					.34874
MARRIED	-.0282	.0506	-.557	.5776					.75217
KIDS	.0831	.0465	1.786	.0741					.37943
FEMALE					-.9856	.0430	-22.924	.0000	.48405
RHO(1,2)	.0572	.0275	2.080	.0375					
	Decomposition of Partial Effect								
	Indirect	Direct	Total						
AGE	.00025	-.00748	-.00723						43.4401
EDUC	-.00061	.01965	.01904						11.3206
INCOME		.06107	.06107						.35208
MARRIED		-.01075	-.01075		(First difference: -.01072)				.75862
KIDS		.03170	.03170		(First difference: .03159)				.40273
FEMALE	.01095		.01095		(First difference: .01096)				.40273

2.13.4 Recursive Bivariate Probit Models

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[y_2 = 1 \mid \mathbf{x}_2] = \Phi(\boldsymbol{\gamma}_2'\mathbf{x}_2),$$
$$\text{Prob}[y_1 = 1 \mid y_2, \mathbf{x}_1, \mathbf{x}_2] = \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1 + \theta y_2, q_{i2}\boldsymbol{\gamma}_2'\mathbf{x}_2, q_{i2}\rho).$$

The dependent variables in the model are

y_1 = presence of a gender economics course,
y_2 = presence of a women's studies program on the campus.

The independent variables in the model are:

z_1 = constant term,
z_2 = academic reputation of the college, coded 1 (best), 2, . . . to 141,
z_3 = size of the full-time economics faculty, a count,
z_4 = percentage of the economics faculty that are women, proportion (0 to 1),
z_5 = religious affiliation of the college, 0 = no, 1 = yes,
z_6 = percentage of the college faculty that are women, proportion (0 to 1),
$z_7 - z_{10}$ = regional dummy variables, South, Midwest, Northeast,West.

The regressor vectors are

$$\mathbf{x}_1 = (z_1, z_2, z_3, z_4, z_5)', \mathbf{x}_2 = (z_2, z_6, z_5, z_7 - z_{10})'.$$

This model is qualitatively different from the bivariate probit model in that the second dependent variable, y_2, appears on the right-hand side of the first equation. This model is a *recursive,*

simultaneous-equations model. The model appears in Heckman (1978), Maddala (1983, p. 123), Greene (2008a, pp. 823-826 and in a spate of recent applications.

The four joint probabilities are

$$P_{11} = \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1 + \theta y_2, \boldsymbol{\gamma}_2'\mathbf{x}_2, \rho),$$
$$P_{10} = \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1, -\boldsymbol{\gamma}_2'\mathbf{x}_2, -\rho),$$
$$P_{01} = \Phi_2[-(\boldsymbol{\gamma}_1'\mathbf{x}_1 + \theta y_2), \boldsymbol{\gamma}_2'\mathbf{x}_2, -\rho],$$
$$P_{00} = \Phi_2(-\boldsymbol{\gamma}_1'\mathbf{x}_1, -\boldsymbol{\gamma}_2'\mathbf{x}_2, \rho).$$

These terms are exactly those of the bivariate probit model that we obtain just by carrying y_2 in the equation for y_1 with no special attention to its endogeneity. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity.

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the first equation, but there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that y_2 equals one. Because y_2 appears in the first equation, this effect is transmitted back to y_1. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, y_1. The reduced form conditional mean is

$$\begin{aligned} E[y_1 \mid \mathbf{x}_1, \mathbf{x}_2] &= \text{Prob}[y_2 = 1]E[y_1 \mid y_2 = 1, \mathbf{x}_1, \mathbf{x}_2] + \\ &\quad \text{Prob}[y_2 = 0]E[y_1 \mid y_2 = 0, \mathbf{x}_1, \mathbf{x}_2] \\ &= \Phi(\boldsymbol{\gamma}_2'\mathbf{x}_2)\ [\ \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1 + \theta, \boldsymbol{\gamma}_2'\mathbf{x}_2, \quad \rho)/\ \Phi(\boldsymbol{\gamma}_2'\mathbf{x}_2)] + \\ &\quad \Phi(-\boldsymbol{\gamma}_2'\mathbf{x}_2)[\ \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1, \quad -\boldsymbol{\gamma}_2'\mathbf{x}_2, -\rho)/\ \Phi(-\boldsymbol{\gamma}_2'\mathbf{x}_2)] \\ &= \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1 + \theta, \boldsymbol{\gamma}_2'\mathbf{x}_2, \rho) + \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_1, -\boldsymbol{\gamma}_2'\mathbf{x}_2, -\rho). \end{aligned} \tag{2.81}$$

Derivatives can be computed using our earlier results for the bivariate normal cdf. The particular feature of interest here is that there is an indirect and a direct effect on y_1 of any variable that appears in both $\mathbf{x}_1$ and $\mathbf{x}_2$. (The indirect effect is the latter.)

2.13.5 A Sample Selection Model

Consider the model analyzed by Boyes, Hoffman and Lowe (1989),

y_{i1} = 1 if individual i defaults on a loan, 0 otherwise,
y_{i2} = 1 if the individual is granted a loan, 0 otherwise.

Wynand and van Praag (1981) also used this framework to analyze consumer insurance purchases in the first application of the selection methodology in a nonlinear model. Greene (1992) applied the same model to y_1 = default on credit card loans, in which y_2 denotes whether an application for the card was accepted or not. Mohanty (2002) used this model to analyze teen employment in California.) For a given individual, y_1 is not observed unless y_{i2} equals 1. Following the lead of the linear regression case, a natural approach might seem to be to fit the second (selection) equation using a univariate probit model, compute the inverse Mills ratio, λ_i, and add it to the first equation as an additional "control" variable to accommodate the selection effect. [This is the approach used by Wynand and van Praag (1981).] The problems with this control function approach are, first, it is unclear what in the model is being "controlled" and, second, assuming the first model is correct, the appropriate model conditioned on the sample selection, is unlikely to

contain an inverse Mills ratio anywhere in it. That result is specific to the linear model, where it arises as $E[\varepsilon_i|\text{selection}]$. What would seem to be the apparent counterpart for this probit model,

$$\text{Prob}(y_{i1} = 1 \mid \text{selection on } y_{i2} = 1) = \Phi(\boldsymbol{\gamma}_1'\mathbf{x}_{i1} + \theta\lambda_i),$$

is not, in fact, the appropriate conditional mean, or probability. For this particular application, the appropriate conditional probability would be

$$\text{Prob}[y_{i1} = 1 \mid y_{i2} = 1, \mathbf{x}_{i1}, \mathbf{x}_{i2}] = \frac{\Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \rho)}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})}. \tag{2.82}$$

We would use this result to build up the likelihood function for the three observed outcomes, as follows:. The three types of observations in the sample, with their unconditional probabilities are

$$\begin{aligned}
y_{i2} = 0&: \text{ Prob}(y_{i2} = 0 \quad\quad\quad\quad \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = 1 - \Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2}),\\
y_{i1} = 0, y_{i2} = 1&: \text{Prob}(y_{i1} = 0, y_{i2} = 1 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = \Phi_2(-\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, -\rho),\\
y_{i1} = 1, y_{i2} = 1&: \text{Prob}(y_{i1} = 1, y_{i2} = 1 \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}) = \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \rho).
\end{aligned} \tag{2.83}$$

The log-likelihood function is based on these probabilities. For further analysis of the response, note that

$$\text{E}[y_{i1} = 1 \mid y_{i2} = 1, \mathbf{x}_{i1}, \mathbf{x}_{i2}] = \text{Prob}[y_{i1} = 1 \mid y_{i2} = 1, \mathbf{x}_{i1}, \mathbf{x}_{i2}],$$

so the interesting partial effects in the model are the partial derivatives of the conditional probability,

$$\begin{aligned}
\frac{\partial E[y_{i1} \mid y_{i2} = 1, \mathbf{x}_{i1}, \mathbf{x}_{i2}]}{\partial \mathbf{x}_{i1}} &= \frac{1}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \frac{\partial \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \rho)}{\partial \mathbf{x}_{i1}} \\
&= \frac{g_{i1}}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \boldsymbol{\gamma}_1, \\
\frac{\partial E[y_{i1} \mid y_{i2} = 1, \mathbf{x}_{i1}, \mathbf{x}_{i2}]}{\partial \mathbf{x}_{i2}} &= \frac{1}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \frac{\partial \Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \rho)}{\partial \mathbf{x}_{i2}} \\
&= \left(\frac{g_{i2}}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} - \frac{\Phi_2(\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, \boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \rho)}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \frac{\phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})}{\Phi(\boldsymbol{\gamma}_2'\mathbf{x}_{i2})} \right) \boldsymbol{\gamma}_2,
\end{aligned} \tag{2.84}$$

where g_{i1} and g_{i2} are defined in (2.74).

The possibility that choice of Public insurance influences the reported health satisfaction is considered in the sample selection model in Table 2.22. The estimate of ρ is high, -.6981, but not statistically significant. The negative estimate does suggest that unobserved factors that it make it more likely that the individual buys the insurance make it less likely that they would report that they are healthier than average, which seems appropriate.

Table 2.22 Estimated Sample Selection Model

	HEALTHY				PUBLIC				
	3911 Individuals Selected				4483 Individuals				
	LogL = -3998.974								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.0056	.2708	.021	.9834	3.7196	.1784	20.849	.0000	1.0000
AGE	-.0178	.0023	-7.799	.0000	.0005	.0026	.200	.8416	43.4401
EDUC	.0857	.0183	4.677	.0000	-.1811	.0099	-18.367	.0000	11.4181
INCOME	.4236	.1659	2.553	.0107	-1.120	.1486	-7.537	.0000	.34874
MARRIED	-.0245	.0505	-.486	.6269					.75217
KIDS	.0962	.0478	2.012	.0442	-.0146	.0553	-.264	.7920	.37943
RHO(1,2)	-.6981	.4139	-1.687	.0916					

2.14 The Multivariate Probit and Panel Probit Models

In principle, a multivariate probit model would simply extend the bivariate probit model to more than two outcome variables just by adding equations. The resulting equation system, again analogous to the seemingly unrelated regressions model, would be

$$
\begin{aligned}
&y_1^* = \boldsymbol{\gamma}_1'\mathbf{x}_1 + \varepsilon_1, \quad y_1 = 1(y_1^* > 0)\\
&y_2^* = \boldsymbol{\gamma}_2'\mathbf{x}_2 + \varepsilon_2, \; y_2 = 1(y_2^* > 0),\\
&\ldots\\
&y_M^* = \boldsymbol{\gamma}_M'\mathbf{x}_M + \varepsilon_M, \; y_M = 1(y_M^* > 0),\\
&\mathrm{E}[\varepsilon_M \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] = 0,\\
&Var[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] = 1,\\
&Cov[\varepsilon_L, \varepsilon_M \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] = \rho_{LM}.
\end{aligned}
\tag{2.85}
$$

The joint probabilities of the observed events, $[y_{i1}, y_{i2} \ldots, y_{iM} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iM}]$, $i = 1, \ldots, n$ that form the basis for the log-likelihood function are the M-variate normal probabilities,

$$L_i = \Phi_M(q_{i1}\boldsymbol{\gamma}_1'\mathbf{x}_{i1}, q_{i2}\boldsymbol{\gamma}_2'\mathbf{x}_{i2}, \ldots, q_{i1}\boldsymbol{\gamma}_M'\mathbf{x}_{iM}, \mathbf{R}),$$

where

$$q_{im} = 2y_{im} - 1,$$

$$\mathbf{R}_{jm} = q_{ij}q_{im}\rho_{jm}.$$

The practical obstacle to this extension is the evaluation of the M-variate normal integrals and their derivatives. Some progress has been made on using quadrature for trivariate integration, but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. However, given the speed of modern computers, simulation-based integration using the GHK simulator or simulated likelihood methods do allow for estimation of relatively large models.

The multivariate probit model in another form presents a useful extension of the random effects probit model for panel data. If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the *panel probit model*,

$$
\begin{aligned}
&y_{it}^* = \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it}, y_{it} = 1(y_{it}^* > 0),\\
&(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}) \sim \mathrm{N}[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

The Butler and Moffitt (1982) approach for this model (as a random effects model) has proved useful in many applications. But, their underlying assumption that $\text{Cov}[\varepsilon_{it},\varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with the restriction that the coefficient vectors are the same in every period, one can obtain a model with free correlations across periods. [Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and 2008a, Example 23.16), and Cappellari and Jenkins (2006) are applications.] Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous as well.

2.15 Endogenous Sampling and Case Control Studies

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be *choice based***.** In the two studies noted, the dependent variable measured the occurrence of a loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y_i = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1 and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

$$\ln L = \sum_{i=1}^{n} w_i \ln F(q_i \boldsymbol{\gamma}'\mathbf{x}_i),$$

where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ in which $\mathbf{H}$ is the weighted second derivatives matrix and $\mathbf{B}$ is the weighted sum of outer products of the first derivatives. (The weights are not squared in computing $\mathbf{B}$.)

The assumption that the population proportions, ω_0 and ω_1 are known in advance is somewhat optimistic. An alternative approach to the problem of choice based sampling is described by Johnson and Albert (1999, pp. 115-118) for the situation of a *case-control* study. Consider an analysis of the occurrence in a population of death from an uncommon disease such as lung cancer. A random sample of individuals would have to be followed (at potentially great expense) for a long time to observe a sample of "responses" and even so, would produce a low proportion of responders in the sampled group. A *retrospective study* might involve searching patient records at a hospital to identify a group of patients who had died from lung cancer along with a set of covariates. Another set of patient records would serve as the controls. The problem with the analysis that now follows is the same as the one in the previous paragraph. The sample is unlikely to be representative of the population.

Let $S_i = 1$ denote the event that an individual in the population is sampled. (The authors are using the term "population" in a subtly different manner than we have to this point. For the case control study described here, the "population" would be the full set of case histories at the hospital, not the full population of individuals in the country (state, world) that might or might not have the disease. Some further assumptions would be needed to argue that what is learned from the population at that hospital could be extended to the (super?)population outside the hospital.) The probability that $S_i = 1$ depends on y_i. The target model is

$$\text{Prob}(y_i = 1|\mathbf{x}_i) = F(\boldsymbol{\gamma}'\mathbf{x}_i),$$

however, the sample information provides only

$$\text{Prob}(y_i = 1|S_i = 1,\mathbf{x}_i),$$

and these may be very different. By Bayes Theorem,

$$\begin{aligned}\text{Prob}(y_i = 1 \mid S_i = 1, \mathbf{x}_i) &= \frac{\text{Prob}(S_i = 1 \mid y_i = 1, \mathbf{x}_i)\text{Prob}(y_i = 1 \mid \mathbf{x}_i)}{\text{Prob}(S_i = 1 \mid \mathbf{x}_i)} \\ &= \frac{\text{Prob}(S_i = 1 \mid y_i = 1, \mathbf{x}_i)\text{Prob}(y_i = 1 \mid \mathbf{x}_i)}{\text{Prob}(S_i = 1 \mid y_i = 1, \mathbf{x}_i)\text{Prob}(y_i = 1 \mid \mathbf{x}_i) + \text{Prob}(S_i = 1 \mid y_i = 0, \mathbf{x}_i)\text{Prob}(y_i = 0 \mid \mathbf{x}_i)}.\end{aligned}$$

Now, make two crucial assumptions to replace the Manski-Lerman assumption of known population proportions. First, assume that the correct specification is a binary logit model and that $\boldsymbol{\gamma}$ contains a constant term, α. Second, assume that the probability that a responder is sampled, $\text{Prob}(S_i = 1|y_i = 1,\mathbf{x}_i) = \lambda_1$ and that the probability that a nonresponder is sampled is also a constant; $\text{Prob}(S_i = 1|y_i = 0,\mathbf{x}_i) = \lambda_0$. That is, the probability of an observation being selected into the sample is independent of the covariates, $\mathbf{x}_i$, in the model. The two assumptions produce

$$\text{Prob}(S_i=1|\mathbf{x}_i) = \lambda_1 \text{Prob}(y_i = 1|\mathbf{x}_i) + \lambda_0 \text{Prob}(y_i = 0|\mathbf{x}_i).$$

It follows that

$$\begin{aligned}\text{Prob}(y_i = 1 \mid S_i = 1, \mathbf{x}_i) &= \frac{\lambda_1 \text{Prob}(y_i = 1 \mid \mathbf{x}_i)}{\lambda_1 \text{Prob}(y_i = 1 \mid \mathbf{x}_i) + \lambda_0 \text{Prob}(y_i = 0 \mid \mathbf{x}_i)} \\ &= \frac{\lambda_1 \dfrac{\exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}}{\lambda_1 \dfrac{\exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)} + \lambda_0 \dfrac{1}{1 + \exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}} \\ &= \frac{\lambda_1 \exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}{\lambda_1 \exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i) + \lambda_0} \\ &= \frac{(\lambda_1/\lambda_0)\exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}{(\lambda_1/\lambda_0)\exp(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i) + 1} = \frac{\exp((\alpha + \tau) + \boldsymbol{\gamma}'\mathbf{x}_i)}{1 + \exp((\alpha + \tau) + \boldsymbol{\gamma}'\mathbf{x}_i)},\end{aligned}$$

where $\tau = \ln(\lambda_1/\lambda_0)$. Therefore, estimation of the binary logit model by maximum likelihood, ignoring the sampling mechanism, produces the familiar consistent estimator of the slope parameters, but a biased estimator of the constant term. The cost of the weaker assumptions in this instance is that the analyst will be unable to obtain predictions of probabilities or partial effects without the reliable estimator of α. But, the benefit is that inference about the slope parameters, themselves, can proceed in spite of the nonrandom sampling mechanism.

3

An Ordered Choice Model for Social Science Applications

The ordered probit model in its modern form was proposed by McElvey and Zavoina (1969, 1971, 1975) for the analysis of ordered, categorical, nonquantitative choices, outcomes and responses. Their application concerned Congressional preferences on a Medicaid bill. [See, as well, the discussion of Gurland et al. (1960) in Section 4.5 which anticipates some aspects of the the social science application.] Familiar recent examples include bond ratings, discrete opinion surveys such as those on political questions, obesity measures, preferences in consumption, and satisfaction and health status surveys such as those analyzed by Boes and Winkelmann (2006a, 2006b) and other applications mentioned in the introduction. The model is used to describe the data generating process for a random outcome that takes one of a set of discrete, *ordered* outcomes. The health satisfaction or opinion survey provide clear examples.

3.1 A Latent Regression Model for a Continuous Measure

The model platform is an underlying random utility model or latent regression model,

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ i = 1,...,n, \tag{3.1}$$

in which the continuous latent utility or 'measure,' y_i^* is observed in discrete form through a censoring mechansm;

$$\begin{aligned} y_i &= 0 \text{ if } \mu_{-1} < y_i^* \leq \mu_0, \\ &= 1 \text{ if } \mu_0 < y_i^* \leq \mu_1, \\ &= 2 \text{ if } \mu_1 < y_i^* \leq \mu_2 \\ &= \ldots \\ &= J \text{ if } \mu_{J-1} < y_i^* \leq \mu_J. \end{aligned} \tag{3.2}$$

Note, for purposes of this introduction, that we have assumed that neither coefficients nor thresholds differ across individuals. These strong assumptions will be reconsidered and relaxed as the analysis proceeds. The vector $\mathbf{x}_i$ is a set of K covariates that are assumed to be strictly independent of ε_i; $\boldsymbol{\beta}$ is a vector of K parameters that is the object of estimation and inference. The n sample observations are labeled $i = 1,...,n$. Long and Freese (2006, p. 183) caution that one ought to ensure that the model to be considered here really is appropriate for the variable of interest before embarking on the analysis. In their case, the question is whether the measured outcome really is ordered. They cite an application of ordering of occupations. Indeed, it is easy to see the validity of their conclusion; the ranking based on, say, some prestige scale is likely to be completely different from a ranking of the same set of outcomes based on expected income. The interpretation of the ordered outcome as a censoring of an underlying continuously measured preference or other measure will provide a reliable guide as to the appropriateness of the model. The thrust of the model is that the observed outcome is not simply a set of discrete outcomes that by some criterion *can* be ordered; the observed outcome is a monotonic (many to one) transformation of a single continuous outcome that naturally *must* be ordered. The further example that Long and Freese pursue, in which the response variable is one of "Strongly

Disagree," "Disagree," "Agree," and "Strongly Agree" is a clear example of a censoring of a naturally ordered underlying preference scale.

The model contains the unknown marginal utilities, β, as well as J+2 unknown threshold parameters, μ_j, all to be estimated using a sample of n observations, indexed by $i = 1,...,n$. The data consist of the covariates, $\mathbf{x}_i$ and the observed discrete outcome, $y_i = 0,1,...,J$. The assumption of the properties of the "disturbance," ε_i, completes the model specification. The conventional assumptions are that ε_i is a continuous random disturbance with conventional cumulative distribution function (cdf), $F(\varepsilon_i|\mathbf{x}_i) = F(\varepsilon_i)$ with support equal to the real line, and that the density, $f(\varepsilon_i) = F'(\varepsilon_i)$ is likewise defined over the real line. The assumption of the distribution of ε_i includes independence from, or exogeneity of, $\mathbf{x}_i$.

The use of models for ordered outcomes arises in many literatures, as suggested in the introduction. The literatures do have focal points at two centers, social sciences including sociology, political science, economics and psychology and in bioassay, as discussed at length below. A reading of the literature in both places suggests that social scientists are broadly comfortable with the idea of the censoring mechanism as the data generating process behind their samples of, usually, individual observations. Their counterparts in bioassay occasionally express some ambivalence about the underlying regression. In Aitchison and Silvey's (1957) canonical application, there is no clear regression-based data generating process at work; if anything the only stimulus in the model is the passage of time, and there are no "coefficients" or "responses" in the equation. Nonetheless, there is a clear, if not perfect, correspondence between their analysis and the ordered choice model. Snell (1964) in contrast, begins development of his model with "We assume there to be an underlying continuous scale of measurement along which the scale categories represent intervals." Once again, however, the analysis to follow has nothing to do with regression; the model relates to discovery of the threshold values in the presence of an individual "effect." But, the applications in the study clearly apply to continuous preference scales, in one case a taste test and in another an opinion survey with answers *terrible*, *poor*, *fair*, *good*, *excellent*.

The use of the latent regression to represent an underlying preference, or utility scale, and the translation of the utility into a discrete indicator has critics in many quarters. A lengthy discussion of the relevance (or irrelevance) of economics to the formulations appears in Hammermesch (2004). On the question, for example, of "how happy does your income make you?" – the question analyzed at some length by Boes and Winkelmann (2006b – see, esp., pp. 4-5) and illustrated below – Hammermesch asks whether it is meaningful to equate this "happiness" with utility. We will then associate the measured outcomes with the supposed utility. [For example, see Groot and van den Bring (2002, 2003b).] For better or worse, this is the position reached by many of the social science applications where the models of ordered choice are applied. They rest crucially on the notion of the underlying regression and the censoring process that produces the measured outcome. Ferrer-i-Carbonell and Frijters (2004) take the discussion yet another level deeper, and consider the underlying assumptions that must be at work in order to use satisfaction measures to reflect underlying welfare measures. [See, as well, Winkelmann and Winkelmann (1998).]

McCullagh (1980) is widely regarded as a codiscoverer of the ordered choice model. [Curiously, he makes no mention of McElvey and Zavoina (1975).] He states (on page 109)

> Motivation for the proposed models is provided by appeal to the existence of an underlying continuous and perhaps unobservable random variable. In bioassay this latent variable usually corresponds to a "tolerance" which is assumed to have a continuous distribution in the population. Tolerances, themselves, are not directly observable but increasing tolerance as manifest through an increase in the probability of survival. The categories are envisaged as contiguous intervals on the continuous scale.... Ordinality is

> therefore an integral feature of such models and the imposition of an arbitrary scoring system for the categories is thereby avoided.

At least to some extent, Anderson and Philips (1981, p. 22) seem unpersuaded;

> It is often possible to argue that an ordered categorical variable is a coarsely measured version of a continuous variable not itself observable. Thus, it is reasonable to assume that the ordered categories correspond to non-overlapping and exhaustive intervals of the real line. ... Although the existence of a latent continuous variable is not crucial for our arguments, it makes interpretation easier and clearer.

They do suggest that in at least one application, a method of predicting the values of the unobservable variable will be developed. Nonetheless, the development of their model begins (on p. 23) with

> Suppose that individuals are grouped into k ordered groups which are identified by an ordered categorical variable y with arbitrarily assigned value s for the sth ordered group; $s = 1,...,k$. The variable y is a convenient identifier for some of the arguments presented later. *The ordering of the groups is not, in general, based on any numerical measurement*. (Emphasis added.)

Anderson (1984, p. 1) in something of a tour de force on ordered outcomes, seems to move in both directions at once:

> Particular emphasis is placed on the case where y is an ordered categorical variable and the category with $y = y_i$ is taken to be "lower" than the category with $y = y_j$ if $i < j$. ... In principle, there is a single unobservable, continuous variable related to this ordered scale, but in practice, the doctor making the assessment will use several pieces of information in making his judgment on the observed category.

The notions of the latent continuous variable and the existence of the latent regression are not mere semantics. At least this is the point behind some of the preceding discussion. Superficially, the same model will arise in any case. However, the underlying platform turns out to be a crucial element of making sense of parameters that are estimated, and of interpretations of the empirical model once obtained from the data. Consider, for example, also from Anderson (1984, p.2).

> The dimensionality of the regression relationship between y and $\mathbf{x}$ is determined by the number of linear functions required to describe the relationship. If only one linear function is required, the relationship is one dimensional; otherwise it is multidimensional. For example, in predicting k categories of pain relief from predictors $\mathbf{x}$, suppose that different functions $\boldsymbol{\beta}_1'\mathbf{x}$ and $\boldsymbol{\beta}_2'\mathbf{x}$ are required to distinguish between the pairs of categories (*worse, same*) and (*same, better*), respectively. *Then the relationship is neither one-dimensional nor ordered with respect to* $\boldsymbol{x}$. (Emphasis added.)

Essentially, the observation is about curve fitting and functional form. One might ask in this instance, "what are the coefficients?" For the current purpose, however, the question would seem to be "what if the simple regression model seems to be inadequate in terms of predicting (by an as yet unspecified procedure) the outcome?" However, the observation raises a vexing question. What if the outcomes, themselves, *are* manifestly ordered. Precisely what does the last sentence imply about the model that is generalized in such a way as to purposely be adequate to handle the full dimensionality of the outcome, as if it were not ordered at all? We will return to this issue below in the context of one of the "generalized" ordered choice models.

3.2 Ordered Choice as an Outcome of Utility Maximization

The appearance of the ordered choice model in the transportation literature falls somewhere between a latent regression approach and a more formal discrete choice interpretation. Bhat and Pulugurta (1998) discuss a model for 'ownership propensity,'

$$C_i = k \text{ if and only if } \psi_{k-1} < C_i^* \leq \psi_k, k = 0,1,...,K, \psi_{-1} = -\infty, \psi_K = +\infty, \quad (3.3)$$

where C_i^* represents the latent auto ownership propensity of household i. The observable counterpart to C_i^* is C_i, typically the number of vehicles owned. [See, e.g., Hensher, Smith, Milthorpe and Bernard (1992). Agyemand-Duah and Hall (1997) apply the model to numbers of trips. Bhat (1997) models the number of non-work commute stops with work travel mode choice.] From here, the model can move in several possible directions: A natural platform for the observed number of vehicles owned might seem to be the count data models (e.g., Poisson) detailed in, e.g., Cameron and Trivedi (1998, 2005) or even a choice model defined on a choice set of alternatives, 0,1,2,… [Hensher et al. (1992)].

The Poisson model for C_i would not follow from a model of utility maximization, though it would, perhaps, adequately *describe* the data generating process. However, a looser interpretation of the vehicle ownership count as a reflection of the underlying preference intensity for ownership suggests an ordered choice model as a plausible alternative platform. Bhat and Pulugurta (1998) provide a utility maximization framework that produces an ordered choice model for the observed count. Their model departs from a random utility framework that assigns separate utility values to different states, e.g., zero car ownership vs. some car ownership, less than or equal to one car owned vs. more than one, and so on (presumably up to the maximum observed in the sample). A suitable set of assumptions about the ranking of utilities produces essentially an unordered choice model for the number of vehicles. A further set of assumptions about the parameterization of the model makes it consistent with the latent regression model above. [See Bhat and Pulugurta (1998, page 64).] A wide literature in this area includes applications by Kitamura (1987, 1988), Golub and van Wissen (1988), Kitamura and Bunch (1989), Golob (1990), Bhat and Koppelmann (1993), Bhat (1996), Agyemara-Duan and Hall (1997), Bhat and Pulugurta (1998) and Bhat, Carini and Misra (1999).

One might question the strict ordering of the vehicle count. For example, the vehicles might include different mixtures of cars, SUVs and trucks. Though a somewhat fuzzy ordering might still seem natural, several authors have opted instead, to replace the ordered choice model with an unordered choice framework, the multinomial logit model and variants. [See, again, Bhat and Pulugurta (1998) who suggest a different utility function for each observed level of vehicle ownership. Applications include Bhat and Pulugurta (1998), Mannering and Winsten (1985), Train (1986), Bunch and Kitamura (1990), Hensher, et al. (1992), Purvis (1994) and Agostino, Bhat and Pas (1996). Groot and van den Brink (2003a) encounter precisely the same issue in their analysis of job training sessions. A count model for sessions seems natural, however the length and depth of sessions differs enough to suggest a simple count model will distort the underlying variable of interest, 'training.'

While many applications appear on first consideration to have some 'natural' ordering, this is not necessarily the case when one recognizes that the ordering must have some meaning also in utility or satisfaction space (i.e., a naturally ordered underlying preference scale) if it assumed that the models are essentially driven by the behavioural rule of utility maximization. The number of cars owned is a good example: 0,1,2, >2 is a natural ordering in physical vehicle space, but it is not necessarily so in utility space.

Ordered and unordered discrete outcome models have distinct conceptual and econometric properties. An unordered model specification is more appropriate when the set of alternative

outcomes representing the dependent variable does not follow a natural ordinal ranking. In unordered models, the utility functions specified by the researcher may not be the same for each alternative. Different attributes may enter into one or more utility expressions, with a general constraint that no single attribute can appear in all utility expressions simultaneously [see Hensher et al. (2005)]. By contrast, the ordered choice model has a single utility expression with thresholds, such as in our example in the introduction.

The discussion to follow will focus on applications in which the underlying choice or intensity variable produces a *naturally strictly ordered observable counterpart*, such as a survey statement of the strength of ones preferences. Save for a brief reconsideration in Section 5.1.2, we will not consider unordered choice models further in this review. [See Hensher, Rose and Greene (2005).] The use of the ordered choice model as a framework for analyzing counts, such as of vehicles owned, remains a possibility under the preceding interpretations. However, once again in the interest of brevity, we will not consider this particular application apart from the general analysis of the model.

3.3 The Observed Discrete Outcome

A typical social science application might begin from a measured outcome such as:

"Rate your feelings about the proposed legislation as

0	Strongly disagree
1	Mildly disagree
2	Indifferent
3	Mildly support
4	Strongly support."

The latent regression model would describe an underlying continuous, albeit unobservable, preference for the legislation as y_i^*. The surveyed individual, even if they could, does not provide y_i^*, but rather, a censoring of y_i^* into five different ranges, one of which is closest to their own true preferences. By the laws of probability, the probabilities associated with the observed outcomes are

$$\text{Prob}[y_i = j \mid \mathbf{x}_i] = \text{Prob}[\varepsilon_i \leq \mu_j - \boldsymbol{\beta}'\mathbf{x}_i] - \text{Prob}[\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i], j = 0,1,...,J. \quad (3.4)$$

It is worth noting, as do many other discrete choice models, the 'model' describes probabilities of outcomes. It does not directly describe the relationship between a y_i and the covariates $\mathbf{x}_i$; there is no obvious regression relationship at work between the observed random variable and the covariates. This calls into question the interpretation of $\boldsymbol{\beta}$, an issue to which we will return at several points below. Though y_i is not described by a regression relationship with $\mathbf{x}_i$ – i.e., y_i is merely a label – one might consider examining the binary variables,

m_{ij} = 1 if $y_i = j$ and 0 if not,

or

M_{ij} = 1 if $y_i \leq j$ and 0 if not,

or

M_{ij}' = 1 if $y_i \geq j$ and 0 if not.

The second and third of these as well as m_{i0} can be described by a simple binary choice (probit or logit) model, though these are usually not of interest. However, in general, there is no obvious regression (conditional mean) relationship between the observed dependent variable(s), y_i, and $\mathbf{x}_i$.

Several normalizations are needed to identify the model parameters. First, in order to preserve the positive signs of all of the probabilities, we require $\mu_j > \mu_{j-1}$. Second, if the support is to be the entire real line, then $\mu_{-1} = -\infty$ and $\mu_J = +\infty$. Since the data contain no unconditional information on scaling of the underlying variable – if y_i^* is scaled by any positive value, then scaling the unknown μ_j and $\boldsymbol{\beta}$ by the same value preserves the observed outcomes – an unconditional, free variance parameter, $\text{Var}[\varepsilon_i] = \sigma_\varepsilon^2$, is not identified (estimable). It is convenient to make the identifying restriction σ_ε = a constant, $\overline{\sigma}$. The usual approach to this normalization is to assume that $\text{Var}[\varepsilon_i|\mathbf{x}_i] = 1$ in the probit case and $\pi^2/3$ in the logit model – in either case to eliminate the free structural scaling parameter. (See Section 2.2.3 for this development ofr binary choice models.) Finally, assuming (as we will) that $\mathbf{x}_i$ contains a constant term, we will require $\mu_0 = 0$. (If, with the other normalizations, and with a constant term present, this normalization is not imposed, then adding a constant to μ_0 and the same constant to the intercept term in $\boldsymbol{\beta}$ will leave the probability unchanged.)

We note at this point a minor ambiguity in the received literature. Some treatments omit the overall constant term in $\boldsymbol{\beta}$ and, in turn, omit the now unnecessary normalization $\mu_0 = 0$. The counterpart in these treatments is $\beta_0 = 0$, where β_0 is the overall constant term. In related fashion, some treatments (e.g., the *Stata* and *SAS* software packages) translate the outcome variable to $y_i = 1,2,...,J$, which produces a different count of possible outcomes. We have maintained the formulation above for two reasons. First, most empirical applications in our experience are based on data that actually contain zero as the origin – e.g., the GSOEP data analyzed by Boes and Winkelmann (2006a, 2006b). Second, as we have formulated the model, the familiar binary choice (probit and logit) models are useful parametric special cases that do not require a reformulation of the entire model. This feature is noted elsewhere by some of the authors discussed below.

The standard treatment in the received literature completes the ordered choice model by assuming either a standard normal distribution for ε_i, producing the "ordered probit" model or a standardized logistic distribution (mean zero, variance $\pi^2/3$), which produces the "ordered logit" model. Applications appear to be well divided between the two. A compelling case for one distribution or the other remains to be put forth – historically, a preference for the logistic distribution has been based on mathematical convenience and because of its ready revelation of "odds ratios" in a convenient closed form. [But, see Berkson (1951) who "prefers logits to probits" in a direct response to Finney. Unfortunately, Berkson's arguments will not help to resolve the issue in the setting of this book.] Contemporary software such as *Stata* and *NLOGIT* have automated menus of other distributional choices, for example, the asymmetric Gompertz and extreme value distributions. However the motivation for these distributions is even less persuasive than that for a preference for probits over logits. These two overwhelmingly dominate the received applications; the others seem more than anything else to be gadgets that are straightforward to program in the software. [An exception is Han and Hausman (1986), who present a model in which an ordered extreme value model emerges naturally. A similar example of duration modeling by Formisiano et al. (2001) is described by Simonoff (2003, pp. 435-448).]

3.4 Probabilities and the Log Likelihood

With the full set of normalizations in place, the likelihood function for estimation of the model parameters is based on the implied probabilities,

$$\text{Prob}[y_i = j \mid \mathbf{x}_i] = [F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)] > 0, j = 0,1,...,J. \quad (3.5)$$

Figure 3.1 shows the probabilities for an ordered choice model with three outcomes,

$$\text{Prob}[y_i = 0|\mathbf{x}_i] = F(0 - \boldsymbol{\beta}'\mathbf{x}_i) - F(-\infty - \boldsymbol{\beta}'\mathbf{x}_i) = F(-\boldsymbol{\beta}'\mathbf{x}_i)$$

$$\text{Prob}[y_i = 1|\mathbf{x}_i] = F(-\boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_1 - \boldsymbol{\beta}'\mathbf{x}_i) \quad (3.4)$$
$$\text{Prob}[y_i = 2|\mathbf{x}_i] = F(-\infty - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_1 - \boldsymbol{\beta}'\mathbf{x}_i) = 1 - F(\mu_1 - \boldsymbol{\beta}'\mathbf{x}_i)$$

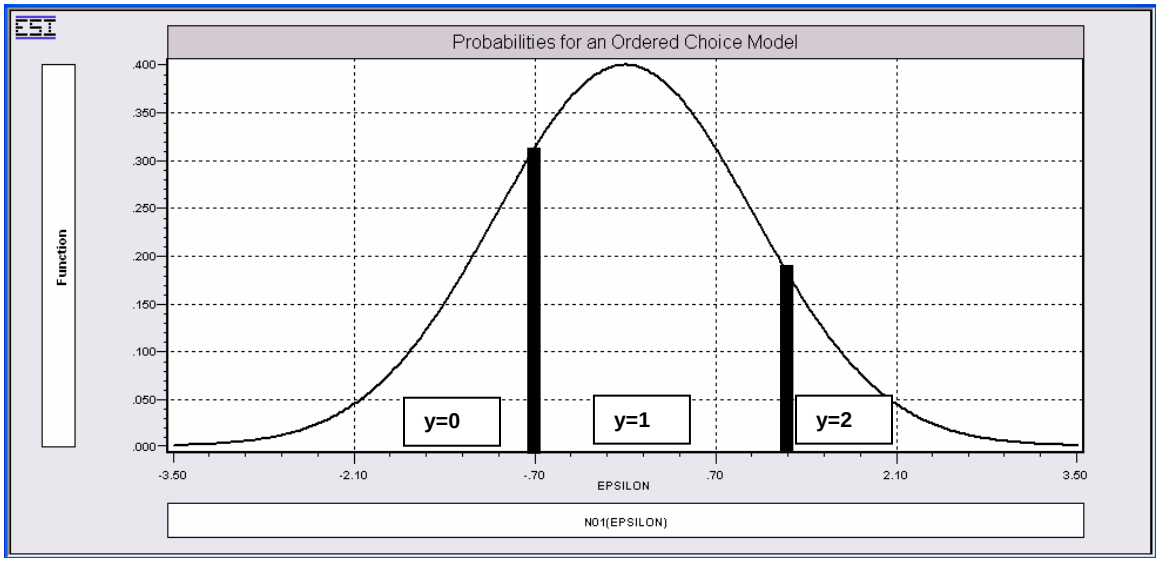


Figure 3.1 Underlying Probabilities for an Ordered Choice Model

3.5 Log Likelihood Function

Estimation of the parameters is a straightforward problem in maximum likelihood estimation. [See, e.g., Pratt (1981) and Greene (2007a, 2008a).] The log likelihood function is

$$\log L = \Sigma_{i=1}^{n}\Sigma_{j=0}^{J} m_{ij} \log[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)], \quad (3.6)$$

where $m_{ij} = 1$ if $y_i = j$ and 0 otherwise. Maximization is done subject to the constraints $\mu_{-1} = -\infty$, $\mu_0 = 0$ and $\mu_J = +\infty$. The remaining constraints, $\mu_{j-1} < \mu_j$ can, in principle, be imposed by a reparameterization in terms of some underlying structural parameters, such as

$$\mu_j = \Sigma_{m=1}^{j} \exp(\alpha_m),$$

however, this is typically unnecessary. (It is necessary in the generalization suggested in Section 8.3 below.) Expressions for the derivatives of the log likelihood can be found in McElvey and Zavoina (1975), Maddala (1983), Long (1997), Stata (2008) and Econometric Software (2007).

The most recent literature (since 2005) includes several applications that use Bayesian methods to analyze ordered choices. Being heavily parametric in nature, they have focused exclusively on the ordered probit model. Some commentary on methods and methodology may be found in Koop and Tobias (2006). Applications to the univariate ordered probit model include Kadam and Lenk (2008), Ando (2006), Zhang et al. (2007) and Tomoyuki and Akira (2006). In the most basic cases, with diffuse priors, the "Bayesian" methods merely reproduce (with some sampling variability) the maximum likelihood estimator. [See Train (2003) for discussion of the Bernstein – von Mises result.] The MCMC methodology is often useful in settings which extend beyond the basic model. We will describe below, for example, applications to a bivariate ordered probit model [Biswas and Das (2002)], a model with autocorrelation [Czado et al. (2005) and

Girard and Parent (2001)] and a model that contains a set of endogenous dummy variables in the latent regression [Munkin and Trivedi (2008).]

3.6 Analysis of Data on Ordered Choices

Analysis of ordered outcomes appears at many points in the literature since its (apparent) emergence with Aitchison and Silvey (1957). As discussed below, what sets McElvey and Zavoina apart is their adaptation to social science applications – the analysis of individual data. The central focus of the applications in bioassay was, and is, on grouped data and the analysis of proportions. The analysis of individual data, in a regression-like setting was relatively new at this point in the literature. Cox (1970), Finney (1971), Theil (1969, 1970, 1971) among others make mention of analysis of individual binary data, but McElvey and Zavoina (1975) seem to be the first the first to extend the ideas of the ordered choice analysis to a model that was closely akin to regression modeling in cross sections of social science data. We will pursue this dichotomy in the next chapter, on the antecedents to the ordered probit (and logit) models.

4

Antecedents and Contemporary Counterparts

McElvey and Zavoina's proposal is preceded by several earlier developments in the statistical literature. The chronology to follow does suggest, however, that their development produced a discrete jump in the received body of techniques. The obvious starting point was the early work on probit methods in toxicology, beginning with Bliss (1934a) and made famous by Finney's (1947b) classic monograph on the subject. The ordered choice model that we are interested in here appears in three clearly discernible steps in the literature, Aitchison and Silvey's (1957) treatment of stages in the life cycle of a certain insect, Snell's (1964) analysis of ordered outcomes (without a regression interpretation) and McElvey and Zavoina's (1975) proposal of the modern form of the "ordered probit regression model." Some later papers, e.g., Anderson (1984) expanded on the basic models.

4.1 The Origin of Probit Analysis: Bliss (1934), Finney (1947)

Bliss (1934a) tabulated graphically the results of a laboratory study of the effectiveness of an insecticide. He plotted the relationship between the "Percent of Aphids Killed" on the ordinate and "Milligrams of Nicotine Per 100 ML of Spray" on the abscissa of a simple figure, reproduced here as Figure 4.1. The figure loosely traces out the familiar sigmoid shape of the normal cdf, and in a natural fashion provides data on what kill rate can be expected for a given concentration of nicotine.

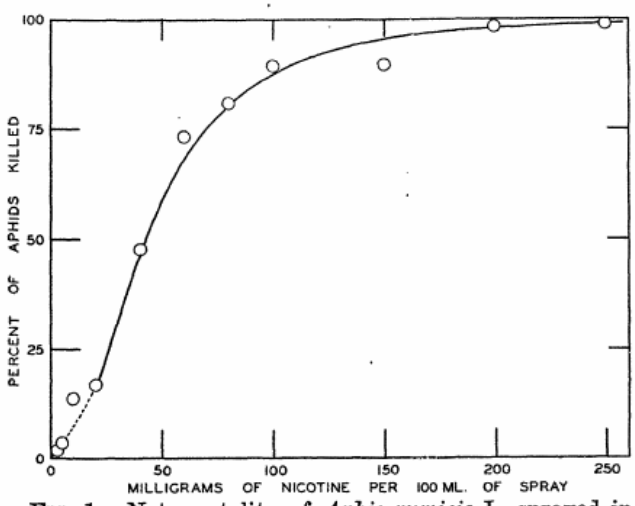


FIG. 1. Net mortality of *Aphis rumicis* L. sprayed in laboratory with different solutions of nicotine; summary of results over 3-year period. Tattersfield and Gimingham.[4] Heavy curve is same as that in Fig. 2 transposed back to original units.

Figure 4.1 Insecticide Experiment

The inverse question – "what concentration is necessary to achieve a given kill rate?" – is answered by inverting the function in the figure. Writing

$$p_i = F(c_i) \tag{4.1}$$

for the former, Bliss suggested that the latter could be answered by analyzing

$$c_i = F^{-1}(p_i). \tag{4.2}$$

The "Method of Probits" is carried out simply by referring the percent kill, p_i to a table to determine the value of c_i of interest. The question can also be answered from the figure by moving eastward from the kill rate of interest to the figure then downward to the concentration. A common application involved elicitation of the lethal dose needed to achieve a 50% kill rate, denoted *LD50*. [See Finney (1944a,b1947a) or (1971), for examples.]

The obvious flaw in the method just described (by the authors, not by Bliss) is that different situations would provide different shaped curves, and the preceding provides no accommodation of that. His search of the then current literature suggested to Bliss that analysts had used a variety of freehand drawing methods to accommodate this kind of heterogeneity, methods that were subject to errors and approximations. Bliss (1934a, p. 38) goes on to suggest "It is believed that these and other difficulties can be minimized if percentage kill and dosage are transformed to units which may be plotted as straight lines on ordinary cross section paper and hence permit fitting by the customary technique of least squares or of the straight line regression equation."

Superficially, Bliss suggests that (4.1) be modified to accommodate the heterogeneity:

$$p_i = F(\alpha + \beta c_i). \tag{4.3}$$

What is needed for the "transformation to units…" is a definition of the specific function, $F(.)$, for which he chose the normal distribution. The inverse transformation is

$$\alpha + \beta c_i = F^{-1}(p_i) = \Phi^{-1}(p_i) = normit(p_i) = y_i. \tag{4.4}$$

It being 1934, computation of the normits was another difficult hurdle. Bliss relied on a table published by Pearson (1914, "Tables of the Normal Probability Integral" in *Pearson's Tables for Statisticians and Biometricians* which is reproduced in Figure 4.2). Dealing with negative numbers was a complication of some substance in 1934, so Bliss suggested the "*probability unit*" or "*probit*"

$$probit(p_i) = normit(p_i) + 5. \tag{4.5}$$

Probits for a number of values of p_i are given in Bliss's Table I reproduced below in Figure 4.2.

These are Bliss's probits. Note that the value associated with 50% is 5.00, not 0.00. A remaining problem is how to handle the extreme tail values. Bliss assigned the value 0.00 to 0.01% and 10.00 to 99.99%. The level of inaccuracy for the intervening values was taken as tolerable. It is intriguing to note, the Pearson Tables (volumes of them) were themselves computed by hand (around 1910). Indeed, though the accuracy of the figures in Bliss's table is noteworthy given when and how they were computed, it is, in fact, quite lacking in absolute terms. Figure 4.3 shows the percentage error in Bliss's (Pearson's) probits (computed using a modern computer and the INP(.) function in *NLOGIT*). It is intriguing to see that the errors are quite large at the tails and clearly not random. An approximation was being used that systematically degrades as the probability moves away from 0.5 in either direction.

TABLE I

Per cent. kill	Probits	Per cent. kill	Probits	Per cent. kill	Probits	Per cent. kill	Probits
1.0	1.87	50.0	5.00	80.0	6.13	95.0	7.21
5.0	2.79	52.0	5.07	81.0	6.18	96.0	7.35
10.0	3.28	54.0	5.14	82.0	6.23	97.0	7.53
15.0	3.61	56.0	5.20	83.0	6.28	98.0	7.76
20.0	3.87	58.0	5.27	84.0	6.34	98.5	7.92
25.0	4.09	60.0	5.34	85.0	6.39	99.0	8.13
30.0	4.30	62.0	5.41	86.0	6.45	99.1	8.18
34.0	4.44	64.0	5.48	87.0	6.51	99.2	8.24
36.0	4.52	66.0	5.56	88.0	6.58	99.3	8.30
38.0	4.59	68.0	5.63	89.0	6.65	99.4	8.38
40.0	4.66	70.0	5.70	90.0	6.72	99.5	8.46
42.0	4.73	72.0	5.78	91.0	6.80	99.6	8.57
44.0	4.80	74.0	5.86	92.0	6.89	99.7	8.69
46.0	4.86	76.0	5.95	93.0	6.98	99.8	8.87
48.0	4.93	78.0	6.04	94.0	7.09	99.9	9.16

Figure 4.2. Table of Probits for Values of p_i.

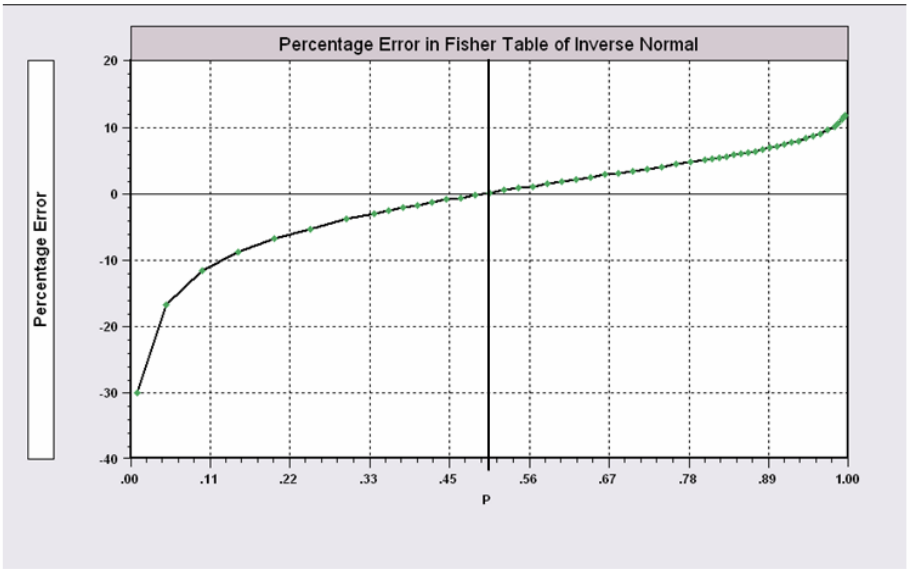


Figure 4.3 Percentage Errors in Pearson Table of Probability Integrals

As the model is stated above, any two points suffice to determine α and β. To accommodate the inevitable sampling variability, the (implied) model must be modified to

$$p_i = \Phi(\alpha + \beta c_i + \varepsilon_i). \quad (4.6)$$

No assumption about the distribution of ε_i is necessary; ε_i is just sampling variability. A mean or median of zero would be a convenient normalization. Bliss then suggests the method of least squares to estimate α and β, which might suggest that he relied (again implicitly) on symmetry of the random errors, ε_i. This would be the evident origin of probit analysis. Other authors had been doing similar analyses for years. But, this was the first point at which the technique was formalized using the inverse probability function (and the normal distribution.) [In Bliss (1934b), the author notes that two other researchers, Hemmingsen (1933) and Gaddum (1933) had used essentially the same method in a study of toxicity in mice.]

Bliss cites several advantages of his method:

(1) It provides a test of normality (of ε). (One could examine the variation of $F^{-1}(p_i)$ around the fitted regression line.)
(2) It includes the ability to do the analysis using logarithms. [See Greene, Knapp and Seaks (1993).] (At least it makes it simpler.)
(3) It suggests a method of determining whether organisms exposed to each dosage were equivalent and the amounts administered experimentally were uniformly proportional to the effective dosage over the range covered by the experiment. (This is examined by exploring the regression relationship.)
(4) It allows the analyst to see "the disclosure of change in the mode of lethal action with certain poisons over different sections of the dosage range indicated by an abrupt change in the slope of the regression." The figure that is shown for this case in the article (shown as Figure 4.4) is equivalent to the introduction of a linear spline in the function based on the log of the dosage, i.e.,

$$p_i = \Phi\{\alpha + \beta \log Dosage_i + \gamma[1(\log Dosage_i > 1.35) \times (\log Dosage_i - 1.35)] + \varepsilon_i\},$$

which is strikingly modern. [See Greene (2008a, pp. 111-112).]
(5) It allows a simple method of expressing in the slope of a straight line, the relative uniformity or diversity between individuals in their susceptibility to a poison. (This seems to relate to the inherent variability of freehand methods used previously.)

In three editions of his celebrated book on the subject of probit analysis, Finney (1947b, 1952, 1971) refined Bliss's methods and applied them to a wide array of experiments. A major practical development in the progression of this work was the advent of software and computers for maximum likelihood methods, including Finney's own contribution to this market, a program that he named BLISS in recognition of his predecessor. [See ISI (1982).]

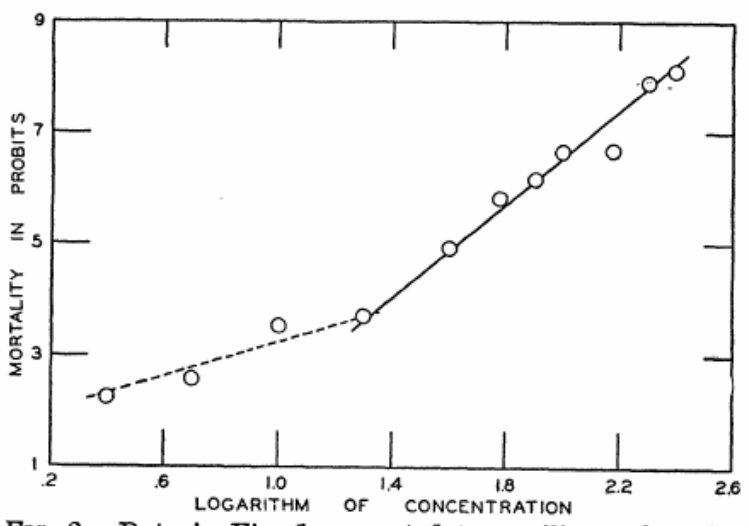


FIG. 2. Data in Fig. 1 converted to rectilinear form by use of logarithms and probits as explained.

Figure 4.4. Implied Spline Regression in Bliss's Probit Model

4.2 Social Science Data and Regression Analysis for Binary Outcomes

To this point, and in the studies noted below, the collection of methods is applied to sampling situations involving grouped data, that is proportions. The samples involved in the analyses described here consisted of observations (n_i,p_i,x_i), $i = 1,...,n$. That is, a group size, a proportion of "responders" and a level of the stimulus. The literature was into the 1970s before researchers began to extend the techniques to individual data. See, for example, the "Frontiers" section of Theil (1971). The formal treatment of individual data for ordered choices – the sort of data observed by social scientists – begins with Walker and Duncan (1967) in the bioassay literature and appeared first in the social sciences in 1971 and 1975 with McElvey and Zavoina's work.

The development of the "minimum chi squared" approach to estimation, and the development of estimation methods as something closer than before to regression analysis might be seen as a bridge between these literatures. Berkson (1944, 1953, 1955a,b, 1957, 1980) and Amemiya (1975, 1980, 1985) suggest an approach to estimation along the lines of

$$p_i = F(\alpha + \beta c_i) + \varepsilon_i. \tag{4.7}$$

[Walker and Duncan (1967), drawing on Gurland and Dahm (1960), also took precisely this approach to modeling probabilities (see p. 169). However, they were concerned with individual data, not sample proportions. We will examine Walker and Duncan's analysis in Section 4.5.] That is, the sampling variability in estimation is laid on the sample proportion, p_i, as an estimator of the population quantity, $F(\alpha + \beta c_i)$. Under this interpretation, the *logit* of p_i, $\log p_i/(1-p_i)$, or *normit transformation*, $\Phi^{-1}(p_i)$ would seem to be less useful, since now the sampling variability is moved inside the function. A two step or iterative application of weighted least squares, the *minimum chi squared* estimator provides an approach that accounts for the nonlinearity of the function and the heteroscedasticity in p_i. [See, e.g., Greene (2003, Section 21.4.6).]

The analysis of the population probability, $F(\alpha + \beta c_i)$, as the conditional mean in a regression relationship can be carried over to a setting of individual data. This line of approach

comes to fruition in the class of "Generalized Linear Models," [Nelder and Wedderburn (1972) and McCullagh and Nelder (1983).] The GLIM approach to modeling binary data embodies the regression interpretation of the probability function and extends easily to the analysis of individual data.

4.3 Analysis of Binary Choice

By 1975, analysis of binary data by social scientists, in grouped or individual form, using maximum likelihood, or minimum chi squared estimators had come to full bloom. The GLIM approach [Grizzle et al. (1969), Nelder and Wedderburn (1972) and Wedderburn (1974), and, see, McCullagh and Nelder (1983) and Pregibon (1984)] had likewise appeared in bioassay. Surveys of varying length of estimation involving binary choices are given in Cox (1970), Finney (1971), Amemiya (1981), Long (1997), Greene (2007a, 2008a) and dozens of other primers and introductions.

4.4 Ordered Outcomes: Aitchison and Silvey (1957), Snell (1964),

Analysis of a dichotomous response (always in grouped form, however), is well developed by the 1940s. Analysis of ordered responses that are of interest in this study, begins in 1957 with an extension to Finney by Aitchison and Silvey (1957). The other relevant antecedent is Snell's (1964) parallel development of an (only apparently) different treatment of ordered outcomes. In what follows, we will use the authors own notation, though contemporary treatments use a uniformly different flavor of notation.

The modeling exercise considered by Aitchison and Silvey (1957) is as follows: Sample observations are made on a species of insect *Petrobius Leash* (*Thysanura, Machilidae*) that passes through s+1 stages in its life cycle. A particular insect is necessarily observed in one stage at any point in time. The last stage is always reached. Observations are made at m different times, denoted x_α, $\alpha = 1,\ldots,m$. The amount of time spent by an insect in stage i, (i=1,…,s) is an observation on a nonnegative random variable, ξ_i. Interest is in estimation of $\lambda_i = E[\xi_i]$ = the average amount of time that will be spent in stage i. The total time spent in stages 1,…,r is $\eta_r = \Sigma_{i=1}^{r}\xi_i$, also a nonnegative random variable. Interest might be in estimation of $\mu_r = E[\eta_r]$ as well. Since $\lambda_i = \mu_i - \mu_{i-1}$, λ_i is estimable from μ_i.

Total time spent in stages up to the observation, η_r, is a continuous random variable with $\Pr(\eta_r \leq x) = G_r(x)$. Probabilities of observation of an insect in the s+1 stages at time x are

$$\begin{aligned}
\pi_1(x) &= \Pr(\eta_1 > x) = 1 - G_1(x), \\
\pi_2(x) &= \Pr(\eta_1 \leq x \text{ and } \eta_2 > x) \\
&= \Pr(\eta_1 \leq x) - \Pr(\eta_1 \leq x \text{ and } \eta_2 \leq x) \\
&= \Pr(\eta_1 \leq x) - \Pr(\eta_2 \leq x) \\
&= G_1(x) - G_2(x).
\end{aligned}$$

(This makes use of the result that if $\eta_r < x$, then $\eta_{r-1} < x$.)

$$\begin{aligned}
\pi_s(x) &= G_s(x) - G_{s-1}(x), \\
\pi_{s+1}(x) &= G_s(x).
\end{aligned}$$

The proportions of insects (subjects) observed in stage s at time x, $p_s(x)$ are moment estimators of $\pi_s(x)$. Estimation of the means is based on the model assumption that the random variables η_r are normally distributed with mean μ_r and standard deviation θ_r, so

$$G_r(x) \quad = \Phi[(x - \mu_r)/\theta_r].$$

The authors consider method of moments estimation of μ_r and θ_r. Let $p_{\alpha r}$ denote the sample estimate of $\pi_r(x_\alpha)$. That is, $p_{\alpha r}$ is the proportion of subjects in stage r at time x_α. Then the relationship above suggests

$$\Phi^{-1}(p_{\alpha r}) = Y_{\alpha r} = x_\alpha/\theta_r - \mu_r/\theta_r.$$

They observe, then, "for given r a straight line fitted to the points $(x_\alpha, Y_{\alpha r})$ will cross the x-axis near the maximum likelihood estimate of μ_r, while the gradient will approximate to the maximum-likelihood estimate of $-\theta_r^{-1}$." By this device, all the parameters of this model may be estimated. Some obvious problems will arise with data sets in which $p_{\alpha r}$ is near zero or one. Moreover, estimation of the scale parameters was complicated (this being 1955), so they considered model simplifications, arriving at $\theta_r^2 = \sigma^2\mu_r$ and then using, instead, maximum likelihood based on the method of scoring. The authors, noting the connection to Finney's work, label this a "generalized probit model." Although the preceding does not involve the same sort of estimation problem as Finney's (in short, the coefficient on x in this model is $1/\theta_r$ and we are, in principle, only estimating the threshold values), there is an obvious relationship. They state (p. 139)

> Clearly a situation might arise where in place of a simple dichotomy, [Finney's case] subjects are divided into more than two classes by any dose of the stimulus. Accordingly, we envisage an experiment where random samples of subjects are subjected to m does x_α ($\alpha = 1,2,\ldots,m$) of a stimulus and as a result of the application of the dose x_α each subject is placed in one of s+1 classes. A straightforward illustration of such an experiment is given by Tattersfield, Gimingham and Morris (1925) who classified insects subject to a poison as *unaffected*, *slightly affected*, *moribund* or *dead*. The particular problem discussed above is another illustration if, in this case, time is regarded as the stimulus.

Thus, Aitchison and Silvey have clearly laid the foundation for the ordered probit model as we now understand it, albeit, the application described does not resemble it very closely at all. They go on to suggest conditions that "must be satisfied in this general experiment in order that the method of analysis used in our particular case should be applicable"

> (i) The classes must be ordered, mutually exclusive and exhaustive.
> (ii) The reactions of a subject to increasing doses must be systematic in the sense that if dose x places a subject in the ith class, then a dose greater than x is required to place this subject in the jth class where j is greater than i.

Point (i) is obvious – the model is designed for ordered outcomes. The second point seems to relate to the latent regression interpretation of the modern view of the model. The authors discuss a "tolerance" for the given classes defined in the model, which the surrounding discussion associates with levels of a latent variable that is observed by the analyst only through the class observed. Finally, the authors note that "if $s = 1$ then the present analysis becomes an ordinary probit analysis and it is in this sense that we have generalized probit analysis."

Before leaving Aitcheson and Silvey, it is interesting to note that although their application did not actually generalize probit analysis, the speculation in the paragraph noted above, in fact, did. The application that they pursued is extended by Feinberg (1980) in what he calls the *continuation ratio model*. [See, as well, Long and Freese (2006, pp. 221-222).] The model is a regression style model that is designed for sequential (so, by implication, ordered) outcomes. The example given by Long and Freese is faculty rank, which would typically include

assistant, then associate, then full professor (and perhaps instructor at the left and chaired professor at the right). The functional form is written for m stages in the progression in which the probability that an observed individual is in stage m given x is $\Pr(y = m|\mathbf{x})$ and the probability that they are in a higher stage is $\Pr(y > m|\mathbf{x})$. Then, the "continuation model" for the log odds is

$$\log\left[\frac{\Pr(y = m \mid x)}{\Pr(y > m \mid x)}\right] = \theta_m - \boldsymbol{\beta}'\mathbf{x}\,.$$

It is not obvious how the ordering aspect of the outcomes enters this model. The requirement in the model (and in university life) that for a given individual,

$$\Pr(y \leq m|\mathbf{x}) \;<\; \Pr(y \leq m+1|\mathbf{x}),$$

is induced by the fact that m+1 means there are more ranks at or below m than m+1, not that the next rank has a higher order than the previous one. For the scenario described, the flaw in the model would seem to be that it is a static model being used to describe a dynamic phenomenon. Although one must pass through the stages in order (though individuals have been known to skip stages), the probabilities in the model do not have any intrinsic relationship to the ordering of the stages, but rather arise the same way if we merely count ranks.

Snell (1964) considers specifically analyzing a set of scores for a ranked set of outcomes such as *Excellent*, *Very Good*, *Good*, *Not Very Good*, *Poor*, *Very Poor*, recorded, perhaps, 6,5,4,3,2,1 or the like. Conventional analysis of such data (Aitchison and Silvey (1957) notwithstanding) was done using analysis of variance techniques, e.g., regression methods assuming (a) normally distributed disturbances and (b) homogeneous variances.

Their departure point is "[w]e assume there to be an underlying continuous scale of measurement along which the scale categories represent intervals." The scale is divided into intervals labeled $k = 0,1,\ldots,k$ by k+2 points, $x_{-1}, x_0, x_1,\ldots,x_k$. Observations in the data, indexed by i, consist of group size, n_i and proportions, p_{ij}, $j = 0,1,\ldots,k$. The underlying continuous distribution function is denoted $P_i(x_j)$. It is unclear what continuous random outcome this is meant to refer to, in connection to the "*i*." However, it is clear from the context that in fact what is implied is that we describe the realization of a random variable, X_i which is the unobserved aforementioned "measurement." Thus, by the construction above, the probability of observing an individual in group i will be in category s_j is equal to

$$P_i(x_j) - P_i(x_{j-1}),\ i = 1,\ldots,m;\ j = 0,\ldots,k.$$

Once again, the reference to "*i*" above refers to a group, so it can only be inferred that what the author has in mind is that group "*i*" consists of n_i realizations of X_i, and the preceding gives the probabilities associated with each member of the group. (Note that there is nothing so far in the data other than the observation subscript, i, to distinguish the groups, e.g., no stimulus x_i.) To continue, "We take the distribution function to be of the form"

$$P_i(x_j) = \;[1 + \exp(-f_{ij})]^{-1} \;=\; \Lambda(f_{ij})\;\text{ (using a contemporary notation).}$$

Finally, f_{ij} is defined to be the "logit" of the "proportion" $P_i(x_j)$,

$$f_{ij} \;=\; \log[P_i(x_j)/(1- P_i(x_j))] = a_i + b_i x_j.$$

The model now has for each i, a location parameter a_i and a spread parameter b_i. To impose homoscedasticity on the data, they assume $b_i = 1$. The log likelihood for the observed data is

$$\log L(a_1,\ldots,a_m,x_{-1},x_0,\ldots,x_k) = \Sigma_{i=1}^{m} n_i \Sigma_{j=0}^{k} p_{ij} \log\left[P_{ij} - P_{i,j-1} \right].$$

It is apparent that a normalization is required to use the entire real line, so $x_0 = -\infty$ and $x_k = +\infty$. He also notes "since the choice of origin is arbitrary, we take $x_1 = 0$." (In fact, since there is no other invariant constant term in the model, this last normalization is not necessary – it now constitutes a substantive restriction.) The remainder of the analysis focuses on methods of estimating m fixed effects a_i and k-2 threshold values, x_j.

The parameters of the model can be loosely estimated by a method of moments type of calculation. Approximate estimates of the threshold values x_k are based on group size weighted averages of the group proportions. Initial estimates of the fixed effects are computed using

$$a_i = -\Sigma_{j=1}^{k} p_{ij} s_j$$

where $s_j = (x_j - x_{j-1})/2$, $j = 2,3,\ldots,k$-1. The two end points corresponding to the lower and upper tails are problematic, and a solution, ultimately, $s_1 = x_1 - 1$ and $s_k = x_{k-1}+1$, is suggested. Newton's method is used to complete the estimation.

Snell's model is functionally equivalent to $P_i(x_j) = \Lambda[x_j - (-a_i)]$ so that the log likelihood function is

$$\log L(a_1,\ldots,a_m,x_{-1},x_0,\ldots,x_k) = \Sigma_{i=1}^{m} n_i \Sigma_{j=0}^{k} p_{ij} \log\left[\Lambda(x_j + a_i) - \Lambda(x_{j-1} + a_i) \right].$$

This corresponds to a modern form of the ordered choice model, though it should be noted that the assumption of a different "effect," a_i for each cross section observation does not appear in the recent literature. (It is estimable, perhaps counter to intuition, because there is more than a single observation for each i; there is a whole set of p_{ij}s for each i.)

It is worth noting as well, that the terms in the log likelihood function above are only positive if the x_j terms are strictly ordered. The initial, "approximate" values will certainly be, because they are functions of the cumulative group proportions. But, the application of Newton's method that follows makes no mention of this restriction, and could break down numerically. The method was only suggested in the text; the author used the approximate, method of moments estimators in the applications.

Some of the closing remarks in the paper are intriguing.

> "The aim throughout this paper has been to present a method based upon a theoretical model and yet to keep the procedure as simple as possible. For this reason, attention has been directed very much towards an approximate solution."

The method of solution is the method of moments; in principle it could have been done with a hand calculator. In 1964, Texas Instruments had just begun production of their first four function calculators, so that might have been optimistic. However, IBMs 7090 series of mainframe computers was already well established and the 360 series was on the near horizon. There would have been no shortage of computing power. A computing language, Fortran (Formula Translation), had been invented in the 1950s. Snell does note that the iterative method "can easily be carried out on a desk machine, and one iteration should be sufficient."

> "The model upon which the method is based takes no account of the experimental design behind the data."

We read this to state that there is no data generating process assumed to be at work here (though, in fact, there must be one in the background – the data arise through some kind of stochastic process; we have attached probabilities to the outcomes.) In fact, the method is semiparametric – the fixed effects approach does stop short of regression. However, the choice of logistic distribution was not entirely innocent. It was made for mathematical convenience, however the numerical results depend on it. The same set of computations could have been done, at considerable cost in complexity, using the normal distribution.

> "Finally, there is no reason why the use of this technique should be restricted to subjective measurement."

Indeed, the recent history has demonstrated the versatility of the method.

4.5 Minimum Chi Squared Estimation of an Ordered Response Model: Gurland, Lee and Dahm (1960)

Gurland, Lee and Dahm (1960) considered the following analysis in bioassay (p. 383): [We will modify their notation slightly so that their model will fit more neatly into the discussion used herein.]

> Suppose N groups consisting of $n_1,\ldots,n_N$ houseflies are exposed to dosages $x_1,\ldots,x_N$, respectively. Out of the n_i flies exposed at dosage x_i, suppose that at the given time of observation,
>
> r_{i1} are dead, r_{i2} are moribund, r_{i3} are alive.
>
> Write the observed proportions as
>
> $$p_{i1} = r_{i1}/n_i,\ p_{i2} = r_{i2}/n_i,\ p_{i3} = r_{i3}/n_i = 1 - p_{i1} - p_{i2}.$$
>
> Let
>
> $$P_{i1} = E[p_{i1}],\ P_{i2} = E[p_{i2}],\ P_{i3} = 1 - P_{i1} - P_{i2}$$
>
> be the corresponding expected proportions or true probabilities. Then, …
>
> $$P_{i1} = \Phi(\alpha_1 + \beta x_i) \quad (1)$$
> $$P_{i1} + P_{i2} = \Phi(\alpha_2 + \beta x_i),\ i = 1,\ldots, \quad (2)$$
>
> where
>
> $$\beta = 1/\sigma,\ \alpha_1 = -\mu_1/\sigma,\ \alpha_2 = -\mu_2/\sigma.$$
>
> … This assumes a normal tolerance distribution $N[\mu_1,\sigma^2]$ of lethal dosages and a normal tolerance distribution $N[\mu_2,\sigma^2]$ of moribund dosages. Furthermore, $\mu_1 > \mu_2$. Since a fly becomes moribund before it dies, the expression in (2), which is the probability a fly is moribund or dead, must involve the same parameter, β as in (1). If the β were not common, the two curves would cross, but this is obviously not permissible since $P_{i1} + P_{i2} > P_{i1}$.

Note, first, the interpretation of P_{ij} as $E[p_{ij}]$ implies $p_{ij} = P_{ij} + \varepsilon_{ij}$, precisely as in Section 4.2. The authors propose a regression approach to estimation of the model parameters, as opposed to maximum likelihood estimation. They proceed to develop a weighted least squares (minimum chi squared) estimator. Second, presumably, the normal distributions assumed above apply to the distributions of tolerances across individual flies. It follows from their analysis, then, that for any particular housefly, $t = 1,\ldots,n_i$,

$$\begin{aligned} \text{Prob}(dead_{i,t}|x_i) &= \Phi[-\mu_1/\sigma + (1/\sigma)x_i] \\ &= \text{Prob}[T^* \leq (1/\sigma)x_i - \mu_1/\sigma], \end{aligned}$$

$$\begin{aligned} \text{Prob}(dead_{it}|x_i) + \text{Prob}(moribund_{it}|x_i) &= \Phi[-\mu_2/\sigma + (1/\sigma)x_i] \\ &= \text{Prob}[T^* \leq (1/\sigma)x_i - \mu_2/\sigma], \end{aligned}$$

where T^* is the tolerance across flies in the experiment. This would appear to be precisely the model ultimately analyzed by McElvey and Zavoina (1975). There is a loose end in the preceding which makes the model an imperfect precursor, however. The authors have avoided the latent regression – they make no mention of it. They state specifically that there are different tolerance distributions with the same variance but different means. But, they do force the same β to appear in both probabilities, arguing that without this restriction, we will be able, for some dosage, x_i to have the probability of dead or moribund be less than that the probability of dead, which is a contradiction of the axioms of probability. It does follow, however, that there are different *prior* distributions for flies that will die after dosage x_i.and flies that will be moribund – i.e., the different tolerance distributions. Thus, there is an ambiguity in the formulation as to what random variable the assumed normal distributions are meant to describe. By a reasonable construction, for example, we might infer that the distribution describes the observed flies only after the reaction to the dosage.

The ambiguities notwithstanding, Gurland et al. (1960) have laid the platform for analysis of ordered outcomes with something resembling a regression approach. The approach is still, however, focused on the analysis of sample proportions. The minimum chi squared (iterated weighted least squares) estimator that they develop is proposed because it "is simpler to apply."

4.6 Individual Data and Polychotomous Outcomes: Walker and Duncan (1967)

Walker and Duncan (1967) were concerned with the problem of using a large number of covariates to analyze the probabilities of outcomes. The experiment in the study involved four large surveys of individuals who were free of heart disease at entry to their study and who were examined long after for the presence of (1) myocardial infarction (*MI*), (2) angina pectoris (*AP*) and (3) no coronary heart disease ($\overline{CHD}$). After considering whether the first two categories might be unordered or ordered, the authors opted to build a model for the latter. Previous analyses had studied cross-tabulated data based on one or two factors and by age and sex. The use of numerous other factors – the application involved 8 in addition to age and sex – necessitated a different approach.

The three outcome model follows along the lines of Gurland et al. (1960) with two major exceptions. First, the large number of factors compels analysis of the individual data, rather than the sample proportions. Second, though only in passing, they note a natural characterization of the data generating process as "Considered jointly they involve the further assumption that the state of an individual described by the vector **x**, which is sufficient to entail the more severe form MI, is certainly sufficient to entail the less severe form *AP*. *If MI and AP are in reality grades of severity of coronary disease, this assumption will hold at least approximately*. If on the other

hand these are distinct, even though closely related diseases, it is not likely to hold." [Emphasis added.] (p. 173.) Coupled with the assumption of the strict ordering of the outcomes, this does sound like the rudiments of an "underlying regression" interpretation. If so, then the authors' assumption of the logistic distribution as shown below completes the formulation of the ordered logit model. Continuing, "The mathematical reflexion of this assumption is seen in the fact that $P_1 + P_2 \geq P_1$, which holds if and only if the 'slope' coefficient β is identical in (6.1) and (6.2), as is easily shown." (In fact, this is only the case if $\alpha_2 > \alpha_1$. Otherwise, it is neither necessary nor sufficient.)

Their three outcome model (where, as before, we have changed their notation for clarity) is:

$$
\begin{aligned}
z_{i1} &= 1 \text{ if } MI_i \text{ and 0 otherwise,} \\
z_{i2} &= 1 \text{ if } AP_i \text{ and 0 otherwise,} \\
z_{i3} &= 1 \text{ if } \overline{CHD}_i \text{ and 0 otherwise,} \\
P_1 &= E[z_{i1}|\mathbf{x}_i], \\
P_2 &= E[z_{i2}|\mathbf{x}_i], \\
P_3 &= 1 - P_1 - P_2, \\
E[z_{i1}|\mathbf{x}_i] &= P_1 = \Lambda(\alpha_1 + \boldsymbol{\beta}'\mathbf{x}_i), \\
E[z_{i1} + z_{i2}|\mathbf{x}_i] &= P_1 + P_2 = \Lambda(\alpha_2 + \boldsymbol{\beta}'\mathbf{x}_i).
\end{aligned}
$$

To preserve the result $P_1 + P_2 \geq P_1$, it must also be true that $\alpha_2 > \alpha_1$. The implied model structure is

$$
\begin{aligned}
\text{Prob}(MI_i|\mathbf{x}_i) &= \Lambda(\alpha_1 + \boldsymbol{\beta}'\mathbf{x}_i), \\
\text{Prob}(AP_i|\mathbf{x}_i) &= \text{Prob}(\textit{Heart Disease}|\mathbf{x}_i) - \text{Prob}(MI_i|\mathbf{x}_i) = \Lambda(\alpha_2 + \beta'\mathbf{x}_i) - \Lambda(\alpha_1 + \beta'\mathbf{x}_i), \\
\text{Prob}(\overline{CHD}_i|\mathbf{x}_i) &= 1 - \Lambda(\alpha_2 + \beta'\mathbf{x}_i).
\end{aligned}
$$

Walker and Duncan are the first to pursue the analysis of ordered probabilities with individual data. In fact, the latent regression model is not necessary to reach their model formulation; we have superimposed our own interpretation on their model to obtain it. They, in turn, did not appear quite ready to make the assumption. Their model is only consistent with that specification. Indeed, what they have proposed is a mathematical model of a set of probabilities that preserve the supposed (severity) ordering of the first and second outcomes. No appeal to a latent regression is needed. On the other hand, quite clearly, it is a small extension to broaden this model to include the formal ordered probit regression model proposed by McElvey and Zavoina (1975).

4.7 McElvey and Zavoina (1975)

McElvey and Zavoina's (1975) proposed model is described at length above. Based on the preceding very short chronology, it would seem that their model was a significant jump forward, not an increment to the existing machinery. In fact, neither Aitchison and Silvey (1957) nor Snell (1964) proposed anything resembling a regression approach to the analysis of ordered outcomes. There is an obvious hint in this direction at the end of the former, but no direct modification of their proposed model would produce a regression style formulation. Certainly, Walker and Duncan's model can easily be made consistent with the structure of McElvey and Zavoina. But, McElvey and Zavoina were the first to formalize the model in terms of an individual choice setting based on a theory of regression, and to develop an effective iterative method of estimation. Walker and Duncan were in similar territory, but they relied on a weighted least squares procedure and an algorithm based on a Kalman filter [Kalman (1960)] that has not reappeared in the literature. McElvey and Zavoina (1975) and Walker and Duncan (1967) were the also the first analysts to propose using individual data. Their predecessors relied entirely on

grouped data (proportions), essentially on the method of moments (or maximum likelihood in a few cases).

4.8 Developments Since McElvey and Zavoina

As noted earlier, McCullagh (1977, 1979, 1980) is credited with codiscovering the ordered choice model. The proposed model, shown below, is precisely a counterpart to the ordered probit model. However, McCullagh stopped short of hanging the framework on a latent regression. Though he departs from "Motivation for the proposed model is provided by appeal to the existence of an underlying continuous random variable," he goes on to state (page 110)

> All the models advocated in this paper share the property that the categories can be thought of as contiguous intervals on some continuous scale. They differ in their assumptions concerning the distributions of the latent variable (e.g. normality (after suitable transformation), homoscedasticity etc.). It may be objected, in a particular example, that there is no sensible latent variable and that these models are therefore irrelevant or unrealistic. However, the models as introduced in Sections 2.1 and 3.1 make no reference to the existence of such a latent variable and its existence is not required for model interpretation. If such a continuous underlying variable exists, interpretation of the model with reference to this scale is direct and incisive. If no such continuum exists the parameters of the models are still interpretable in terms of the particular categories recorded and not those which might have obtained had the defining criteria $\{\theta_j\}$ been different. Quantitative statements of conclusions are therefore possible in both cases although more succinct and incisive statements are usually possible when direct appeal to a latent variable is acceptable.

McCullagh seems to be holding back from a commitment to an underlying regression. As he notes, however, it will emerge ultimately that interpretation of the coefficients of the model without such an assumption becomes a bit ambiguous.

Though the idea of the ordered logit model shown below is sometimes attributed to McCullagh, elements of it appear earlier in Andrich (1979) and Plackett (1974), and McCullagh cites Plackett for some of his results. The model proposed is based on a discrete random variable with "k ordered categories of the response" with probabilities $\pi_1(\mathbf{x})$, $\pi_2(\mathbf{x})$, ..., $\pi_k(\mathbf{x})$. ("In the case of two groups, $\mathbf{x}$ is an indicator variable or two level factor indicating the appropriate group." This appears to suggest a contingency table sort of analysis, for which the "ordering" would be superfluous.) The response variable, Y, takes values $y = 1,...,k$ with the listed probabilities. Define $\kappa_j(\mathbf{x})$ to be the odds that $Y \leq j$ given $\mathbf{x}$. Then, the "proportional odds model" specifies that

$$\kappa_j(\mathbf{x}) = \kappa_j \times \exp(-\boldsymbol{\beta}'\mathbf{x}), j = 1,...,k.$$

The ratio of corresponding odds is

$$\kappa_j(\mathbf{x}_1)/\kappa_j(\mathbf{x}_2) = \exp[-\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)],$$

which is independent of j and depends only on the difference between the covariate vectors. Given the odds ratio stated as above and defining $\gamma_j(\mathbf{x}) = \pi_1(\mathbf{x}) + ... + \pi_j(\mathbf{x})$, the *proportional odds model* becomes equivalent to

$$\log[\gamma_j(\mathbf{x})/(1-\gamma_j(\mathbf{x}))] = \theta_j - \boldsymbol{\beta}'\mathbf{x}, j = 1,...,k.$$

This is mathematically identical to the familiar ordered choice model discussed earlier. Formally, using a more recent notation,

$$\text{Prob}[y \leq j] = \Lambda(\theta_j - \boldsymbol{\beta}'\mathbf{x}),$$

which is the ordered logit model. As the author notes, no appeal to an underlying regression model is necessary to achieve this result. Remaining to be determined is the mechanism by which the observed discrete random variable is assigned to *k* exhaustive, exclusive and *ordered* categories. The model is meant to apply to proportions, as shown in a series of applications that follows. The application that follows immediately, however, does fall naturally into the latent continuous measure framework, a study of tonsil sizes in a sample of 1,398 children [Holmes and Williams (1954)], shown in Figure 4.5.

TABLE 1
Tonsil size of carriers and non-carriers of Streptococcus pyogenes

	Present but not enlarged	*Enlarged*	*Greatly enlarged*	*Total*
Carriers	19	29	24	72
Non-carriers	497	560	269	1326
Total	516	589	293	1398

Figure 4.5 McCullagh Application of Ordered Outcomes Model

For the simple case shown above, interpretation of the β in the "regression" will be simple, as it will highlight the differences in the probabilities or odds for the outcomes in the two groups. For more complicated kinds of regressors, for example, if age, height, or weight appeared in the data set above, then interpretation of the coefficients would be much more complicated without resort to a regression model of some sort, and a notion of "holding other things constant." In his analysis of this data set, Tutz (1990, 1991) argues that the higher outcomes (more to the right) can only be reached by passing through the lower ones. This calls for a different approach, which he labels the *sequential model*. The simplest case would be Agresti's (1984) *continuation ratio model*,

$$\text{Prob}(y = r \mid y \geq r, \mathbf{x}) = D(\theta_r - \boldsymbol{\beta}'\mathbf{x}),$$

where $D(.)$ is a transformation of the index. This yields the unconditional probabilities

$$\text{Prob}(y = r \mid \mathbf{x}) = D(\theta_r - \boldsymbol{\beta}'\mathbf{x})\prod_{i=1}^{r-1}\left[1 - D(\theta_i - \boldsymbol{\beta}'\mathbf{x})\right].$$

A variety of extensions are suggested. [For another survey of this and related models, see Barnhart and Sampson (1994).]

Anderson and Philips (1981) continue McCullagh's development in two directions. Researchers in this area work back and forth around the assumption of the latent continuous variable and latent regression. Second, they introduced some results related to functional form. As noted earlier, their departure point is "... an ordered categorical variable is a coarsely measured version of a continuous variable not itself observable." The model proposed is as follows: "[I]ndividuals are grouped into *k* ordered groups which are identified by an ordered categorical variable *y* with arbitrarily assigned value *s* for the *s*th ordered group; $s = 1,...,k$. ... The ordering of groups is not, in general, based on any numerical measurement." (The authors are holding back from the assumption. However, one might ask, on what basis *is* the ordering of groups assigned if not some underlying quantitative measure?) A regressor vector, $\mathbf{x}$, is defined. The Plackett (1974, 1981) and McCullagh (1980) functional form is

$$\text{Prob}(y \le s \mid \mathbf{x}) = \frac{\exp(\theta_s - \boldsymbol{\beta}'\mathbf{x})}{1+\exp(\theta_s - \boldsymbol{\beta}'\mathbf{x})}, s = 0,1,...,k,$$

where $\theta_1 \le \theta_2 \le ... \le \theta_{k-1}$, $\theta_0 = -\infty$, $\theta_k = +\infty$. (The author uses weak inequalities, though in order to prevent zero probabilities for non-null events, strong inequalities are required.) It follows, as we observed earlier, that

$$\text{Prob}(y = s|\mathbf{x}) = \Lambda(\theta_s - \boldsymbol{\beta}'\mathbf{x}) - \Lambda(\theta_{s-1} - \boldsymbol{\beta}'x),$$

which is the "logistic model." This is also labeled the "cumulative odds model" by McCullagh (1980). The authors suggest, instead, that we write

$$\text{Prob}(y \le s|\mathbf{x}) = \Psi(\theta_s - \boldsymbol{\beta}'\mathbf{x}),$$

where $\Psi(.)$ is a "completely specified cumulative distribution function." This is a generalized "linear" model, but "nonlinear" versions are possible and are referred to in the discussion. The above models will be called "*ordered regression models*." (Emphasis added. This is the first occurrence of the term that we have encountered in this literature search.)

The authors justify the model in terms of a latent unobservable, z, where, conditioned on $\mathbf{x}$, z has a logistic distribution. Although z is not observed, a related, grouped version of z, y, is observable. Of course, this is precisely the interpretation that McElvey and Zavoina have provided for the model. (Once again, however, there is no mention of McElvey and Zavoina or their model.) We have on the suggested basis,

$$y = s \text{ if } \theta_{s-1} \le z < \theta_s \ (s = 1,...,k).$$

Note that assumptions are made only about the conditional distribution of z given $\mathbf{x}$ and y given $\mathbf{x}$. No assumption is made about the marginal distribution of $\mathbf{x}$, which prompts the claim that these models make only moderate distributional assumptions.

"Other assumptions are possible for the form of the distribution of z given $\mathbf{x}$. One obvious choice is that this should be the normal distribution, $N(\boldsymbol{\beta}'\mathbf{x},1)$, leading to the probit model,

$$\text{Prob}(y \le s|\mathbf{x}) = \Phi(\theta_s - \boldsymbol{\beta}'\mathbf{x}).$$

Here, $\Phi(.)$ represents the usual probit function. For practical purposes, the logistic and probit models are virtually indistinguishable, but the logistic model of (1) and (2) is often preferred for its computational convenience." [Anderson and Philips (1981).] Thus, the ordered probit model is (re)born, here in 1981.

Aitchison and Bennett (1970) is occasionally cited as another antecedent to the ordered choice models considered here. In fact, they were concerned with a different setting altogether, though it is intriguing to note that their formulation is precisely that used to motivate McFadden's conditional logit model (1974). Since they did not consider ordered outcomes, we will forego a detailed discussion of their results.

4.9 Other Related Models

Many authors have modified these models at various edges for different situations and types of data. Some major references to examine for details are Agresti (1984, 1990), Clogg and Shihadeh (1994) and Greenwood and Farewell (1988). Before closing this review, we note two that have particular relevance for our discussion.

4.9.1 Known Thresholds

Stewart (1983), Terza (1985) and Bhat (1994) examine a setting in which essentially the conditions of the ordered probit model emerge, save that there is more information about the censoring than merely the categories. An obvious example considered by these authors is given by bracketed income data. When income data are censored into known ranges, the resulting data generating process is precisely that of the ordered choice model except that the threshold values are known. Suppose, for example, that y^* = log of income is normally distributed with mean $\mu = \boldsymbol{\beta}'\mathbf{x}$ and variance σ^2, so

$$y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon,$$

and the censoring mechanism is

$$y = j \text{ if } A_{j-1} < y^* \leq A_j,$$

where A_{j-1} and A_j are known values. Then, the log likelihood is built up from the probabilities for the observed outcomes;

$$\text{Prob}(y = j|\mathbf{x}) = \left[\Phi\left(\frac{A_j - \boldsymbol{\beta}'\mathbf{x}}{\sigma}\right) - \Phi\left(\frac{A_{j-1} - \boldsymbol{\beta}'\mathbf{x}}{\sigma}\right)\right]. \tag{4.8}$$

For this model, the parameters $\boldsymbol{\beta}$ and σ are both identified (estimable). The ordering of the outcomes is enforced a fortiori by the ordering of the known brackets. This model is, in fact, not a discrete choice model in the spirit of the others that are considered here. Rather, it is a less complicated censoring model more closely resembling the tobit model. [Tobin (1958), Amemiya (1985a, 1985b), Greene (2008a).] There is a temptation to treat this model using linear regression analysis, substituting, e.g., the midpoints of the brackets for intermediate values and some reasonable value for the upper and lower ranges. The temptation should be resisted, since (1) the likelihood for the data and the structural parameters is well defined (and the estimator is available as a preprogrammed procedure in modern software) and (2) least squares in this setting will be inconsistent. The OLS estimator will suffer from truncation bias. The overall result is that because there is variation in $\mathbf{x}$ that is not associated with variation in y, the OLS slopes will tend to be biased toward zero. The maximum likelihood estimator, which does not display this feature, is easily obtained. We do note, however, if, instead of midpoints, one uses for the substituted values

$$E[y^*|\, A_{j-1} < y^* \leq A_j, \mathbf{x}] = \boldsymbol{\beta}'\mathbf{x} + \sigma\left[\frac{\phi[(A_{j-1} - \boldsymbol{\beta}'\mathbf{x})/\sigma] - \phi[(A_j - \boldsymbol{\beta}'\mathbf{x})/\sigma]}{\Phi[(A_j - \boldsymbol{\beta}'\mathbf{x})/\sigma] - \Phi[(A_{j-1} - \boldsymbol{\beta}'\mathbf{x})/\sigma]}\right], \tag{4.9}$$

then, with an appropriate iterate for σ as well as this implicit estimator for $\boldsymbol{\beta}$, this is equivalent to the EM algorithm [see Dempster, Laird and Rubin (1977)], and is an effective, albeit inefficient way to compute the maximum likelihood estimators of σ and $\boldsymbol{\beta}$. (It will be slow to converge compared to other gradient methods such as Newton's method.)

4.9.2 Nonparallel Regressions

A second modification of the model, due to Anderson (1984) is of interest here. He notes (p. 4) "The ordering of the categories, or subsets of them, with respect to the regression variables is open to question in some cases. Hence, we start with the logistic regression model suitable for a qualitative, categorical response variable [Cox (1970), Anderson (1972)]." This is

$$\text{Prob}(y = y_s \mid \mathbf{x}) = \frac{\exp(\beta_{0s} - \boldsymbol{\beta}_s'\mathbf{x})}{\Sigma_{t=1}^{k} \exp(\beta_{0t} - \boldsymbol{\beta}_t'\mathbf{x})},$$

where $\beta_{0k} = 0$ and $\boldsymbol{\beta}_k = \mathbf{0}$ are introduced to simplify the notation. In fact, the function listed is homogeneous of degree zero, and the "simplifications" are normalizations needed for identification. This is precisely the multinomial logit model developed by McFadden, (1974) and Nerlove and Press (1972). Characteristically (apparently), there is no connection across the branches of the literature. (This being before the Internet, perhaps the lack of connection across disparate literatures is an understandable consequence of the difficulty of a detailed search. We take that sort of thing for granted now.) Anderson proposes this model for unordered categorical outcomes. He notes, in passing, however, that this model often "gives a good fit" even when the $\boldsymbol{\beta}$s are "restricted to be parallel." "This is particularly true when the categories are ordered." That is to suggest, the ordered choice model considered thus far embodies the *restriction* that the $\boldsymbol{\beta}$s are the same. By a simple transformation of the ordered logit model, we find

$$\text{logit}(j) = \log[\text{Prob}(y \leq j \mid \mathbf{x}) / \Pr(y > j \mid \mathbf{x})] = \mu_j - \boldsymbol{\beta}'\mathbf{x}, \tag{4.10}$$

which means that $\partial\text{logit}(j)/\partial\mathbf{x} = \boldsymbol{\beta}$ for all j. This has come to be known as the "parallel regressions assumption." [See, e.g., Long (1997, p. 141).] This feature of the model has motivated one form of the "generalized ordered logit" (and probit) model. We will reconsider this generalization of the model in some detail below.

5

Estimation, Inference and Analysis Using the Ordered Choice Model

In this chapter, we will survey the elements of estimation, inference and analysis with the ordered choice model. It will prove useful to develop an application as part of the discussion.

5.1 Application of the Ordered Choice Model to Self Assessed Health Status

Riphahn, Wambach and Million (RWM, 2003) analyzed individual data on health care utilization (doctor visits and hospital visits) using various models for counts. The data set is a large panel extracted from the German Socioeconomic Panel (GSOEP). [See RWM (2003) and Greene (2008a) for discussion of the data set in detail.] The data set is an unbalanced panel including 7,293 German households observed from 1 to 7 times and a total of 27,326 observations. (We will visit the panel data aspects of the data and models later.) Among the several interesting variables in this data set is HSAT, a self reported health assessment that is recorded with values 0,1,..,10 (so, $J = 10$). Figure 5.1 shows the distribution of outcomes for the full sample: The figure reports the variable NewHSAT, not the original variable. Forty of the 27,326 observations on HSAT in the original data were coded with noninteger values between 6.5 and 6.95. We have changed these 40 observations to 7s. In order to construct a compact example that is sufficiently general to illustrate the technique, we will aggregate the categories shown as follows: (0-2)=0, (3-5)=1, (6-8)=2, (9)=3, (10)=4. [One might expect collapsing the data in this fashion to sacrifice some information and, in turn, produce a less efficient estimator of the model parameters. See Murad et al. (2003) for some analysis of this issue.] Figure 5.2 shows the result, once again for the full sample, stratified by gender. The families were observed in 1984-1988, 1991 and 1995. For purposes of the application, to maintain as closely as possible the assumptions of the model, at this point, we have selected the most frequently observed year, 1988, for which there are a total of 4,483 observations, 2,313 males and 2,170 females. We will use the following variables in the regression part of the model,

$$\mathbf{x} = (\textit{constant, Age, Income, Education, Married, Kids}).$$

In the original data set, *Income* is HHNINC (household income) and *Kids* is HHKIDS (household kids). *Married* and *Kids* are binary variables, the latter indicating whether or not there are children in the household. Descriptive statistics for the data used in the application are shown in Table 2.1. We have used the same independent variables with the new ordered dependent variable.

5.2 Distributional Assumptions

As suggested earlier, one of the ambiguities in the set of procedures for ordered choice modeling is the distributional assumption. There seems to be little to determine whether the logit, probit, or some other distribution is to be preferred. The logistic model has some mathematical features to recommend it, but any of these, such as the computation of odds ratios can be replicated under other assumptions, perhaps at some minor inconvenience (depending on one's

software). The deeper question of how the distributional assumption relates to the model structure remains unresolved. Stewart (2003) proposes, beyond the familiar choices a "seminonparametric generalized ordered probit" that is considerably more complicated than the logit and probit models examined here. The model is automated in a *Stata* command however. Stewart's and other semiparametric approaches are developed in Chapter 12. We do note, the offered procedure produces coefficient estimates, but it is unclear how these can be translated into partial effects or other useful quantities. It remains true in this (and all parametric and semiparametric forms) that the vector of partial effects is a scalar multiple of β. On this basis, Stewart argues that ratios of coefficients are useful substitutes for partial effects.

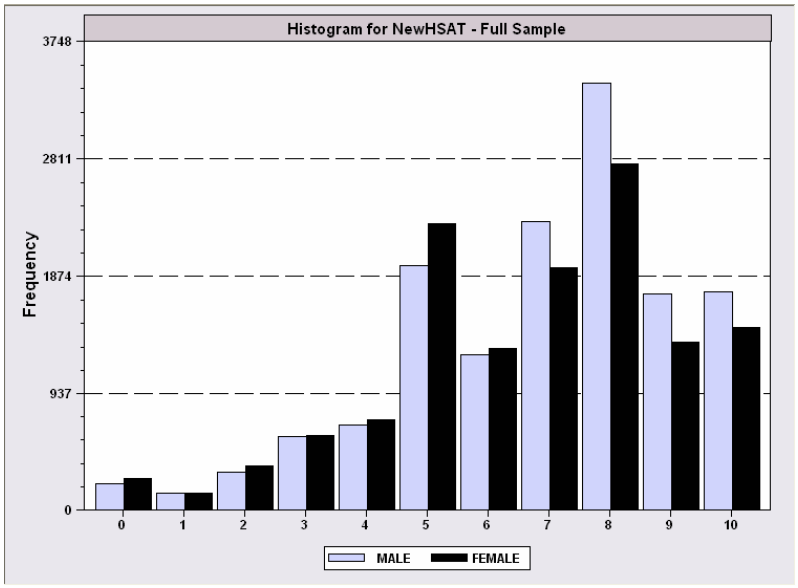


Figure 5.1 Self Reported Health Satisfaction

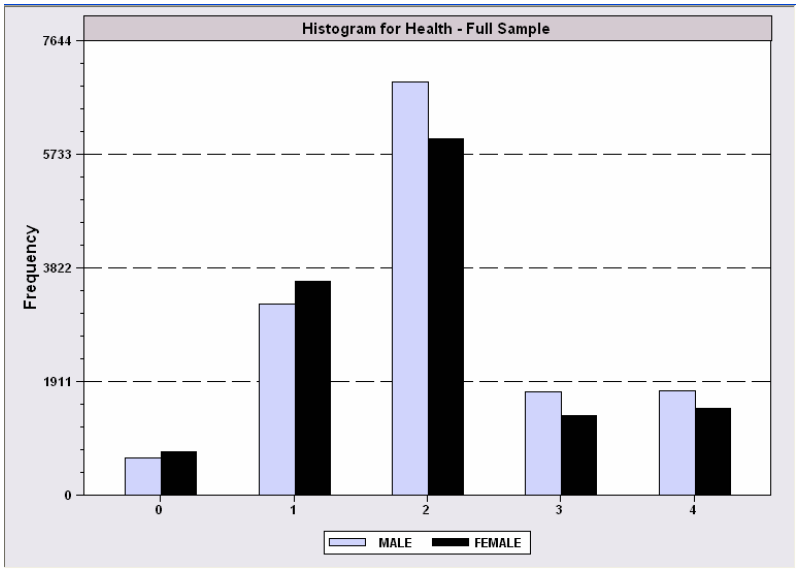


Figure 5.2 Health Satisfaction with Combined Categories

5.3 The Estimated Ordered Probit (Logit) Model

Table 5.1 presents estimates of the ordered probit and logit models for the 1988 data set. (Results from the computer program have been extracted and blended to display the estimates. All computations were carried out using *NLOGIT4*. They can all be replicated with equal convenience with *Stata* and, perhaps with a bit more programming, with *EViews*, *TSP*, *SAS* and most other commercial programs.) The tabulated results include diagnostic statistics such as the log likelihood function, a description of the observed data on the outcome, followed by standard presentations of the coefficients, standard errors, etc. These will be examined in detail in the sections to follow.

The estimates for the probit model imply

$$y^* = 1.97882 - .01806Age + .03556Educ + .25869Income - .03100Married + .06065Kids + \varepsilon.$$

$$y = 0 \text{ if } y^* \leq 0$$
$$y = 1 \text{ if } 0 < y^* \leq 1.14835$$
$$y = 2 \text{ if } 1.14835 < y^* \leq 2.54781$$
$$y = 3 \text{ if } 2.54781 < y^* \leq 3.05639$$
$$y = 4 \text{ if } y^* > 3.05639.$$

Figure 5.3 shows the implied model for a person of average age (43.44 years), education (11.418 years) and income (0.3487) who is married (1) with children (1). The figure shows the implied probability distribution in the population for individuals with these characteristics. As we will examine in the next section, the force of the regression model is that the probabilities change as the characteristics (**x**) change. In terms of the figure, changes in the characteristics induce changes in the placement of the partitions in the distribution and, in turn, in the probabilities of the outcomes.

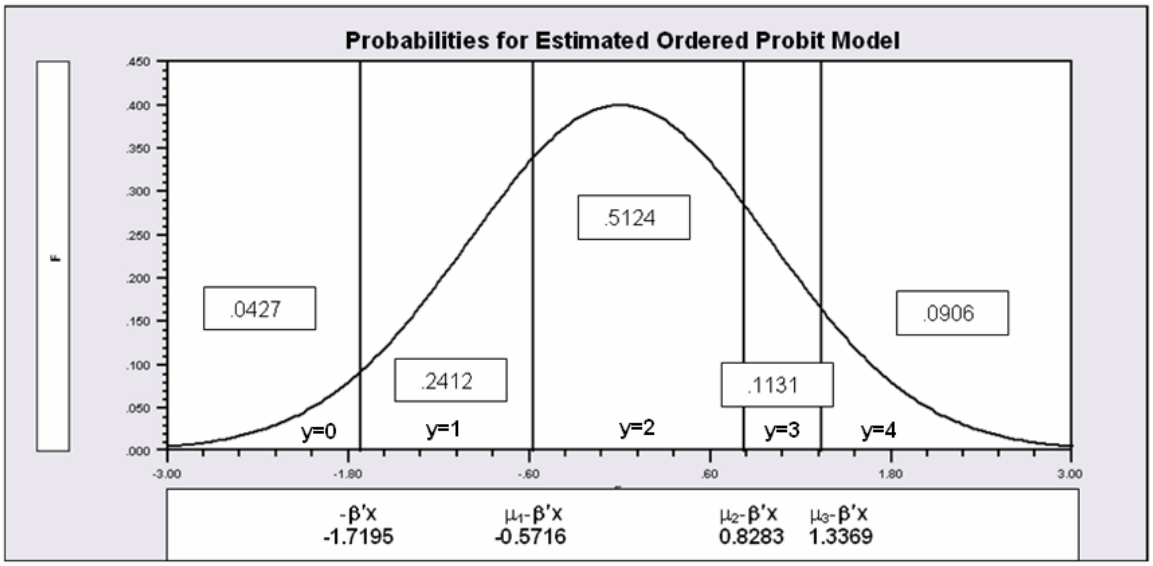


Figure 5.3 Estimated Ordered Probit Model

Table 5.1 Estimated Ordered Choice Models: Probit and Logit

TABLE OF CELL FREQUENCIES FOR ORDERED PROBABILITY MODEL

Outcome	Frequency Count	Frequency Percent	Cumulative < = Count	Cumulative < = Percent	Cumulative > = Count	Cumulative > = Percent
HEALTH=00	230	5.1305	230	5.1305	4483	100.0000
HEALTH=01	1113	24.8271	1343	29.9576	4253	94.8695
HEALTH=02	2226	49.6542	3569	79.6119	3140	70.0424
HEALTH=03	500	11.1532	4069	90.7651	914	20.3881
HEALTH=04	413	9.2349	4483	100.0000	414	9.2349

	Logit				Probit				
	LogL = -5749.157				LogL = -5752.985				
	LogL0 = -5875.096				LogL0 = -5875.096				
	Chisq = 251.8798				Chisq = 244.2238				
	PseudoRsq = .0214362				PseudoRsq = .0207847				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	3.5179	.2038	17.260	.0000	1.9788	.1162	17.034	.0000	1.0000
AGE	-.0321	.0029	-11.178	.0000	-.0181	.0016	-11.166	.0000	43.4401
EDUC	.0645	.0125	5.174	.0000	.0356	.0071	4.986	.0000	11.4181
INCOME	.4263	.1865	2.286	.0223	.2587	.1039	2.490	.0128	.34874
MARRIED	-.0645	.0746	-.865	.3868	-.0310	.0420	-.737	.4608	.75217
KIDS	.1148	.0669	1.717	.0861	.0606	.0382	1.586	.1127	.37943
Mu(1)	2.1213	.0371	57.249	.0000	1.1484	.0212	54.274	.0000	
Mu(2)	4.4346	.0390	113.645	.0000	2.5478	.0216	117.856	.0000	
Mu(3)	5.3771	.0520	103.421	.0000	3.0564	.0267	115.500	.0000	

5.4 The Estimated Threshold Parameters

The sample proportions might provide a motivation to choose the underlying distribution to match the histogram of the observed outcome variable. But, the sample proportions in the ordered choice model do not provide a histogram of the underlying distribution. For example, Figure 5.4a provides a histogram of the variable "Husband's Occupation" according to the Hollingsworth scale (coded 1 – 6) in a sample of 6,366 observations. [See Greene, 2008, Appendix Table F24.1.] The data seem to suggest a leftward skew and might suggest a nonnormal distribution such as the complementary log log model were one to consider an ordered choice model for this variable. However, there is nothing in the formulation that would suggest a nonnormal distribution for the underlying random utility model. The threshold parameters adjust to allocate the mass of the distribution to mimic the sample, For this example, if the model were simply

$$y^* = \alpha + \varepsilon$$
$$y = j \text{ if } \mu_{j-1} < y^* \leq \mu_j, j = 0,1,2,3,4,5,$$

(we have subtracted one from the observed variable), then the only parameters estimated would be the constant term and the four thresholds. The six sample proportions, the sample cumulative proportions, and implied values of the parameters are as follows:

y	0	1	2	3	4	5
p	.0360	.2054	.0770	.3189	.2795	.0833
F	.0360	.2414	.3184	.6373	.9167	1.0000
$\Phi^{-1}(F)$	$-\alpha$	$\mu_1 - \alpha$	$\mu_2 - \alpha$	$\mu_3 - \alpha$	$\mu_4 - \alpha$	
Value	-1.80	-0.70	-0.47	0.35	1.38	$+\infty$

Figure 5.4a,b shows the partitioning of the underlying normal distribution that is consistent with these frequencies. The thresholds will adjust so that the probabilities from the normal distribution will match the sample proportions. Note that this allocation is fully consistent with the underlying normal distribution in spite of the somewhat non-normal appearance of the sample proportions.

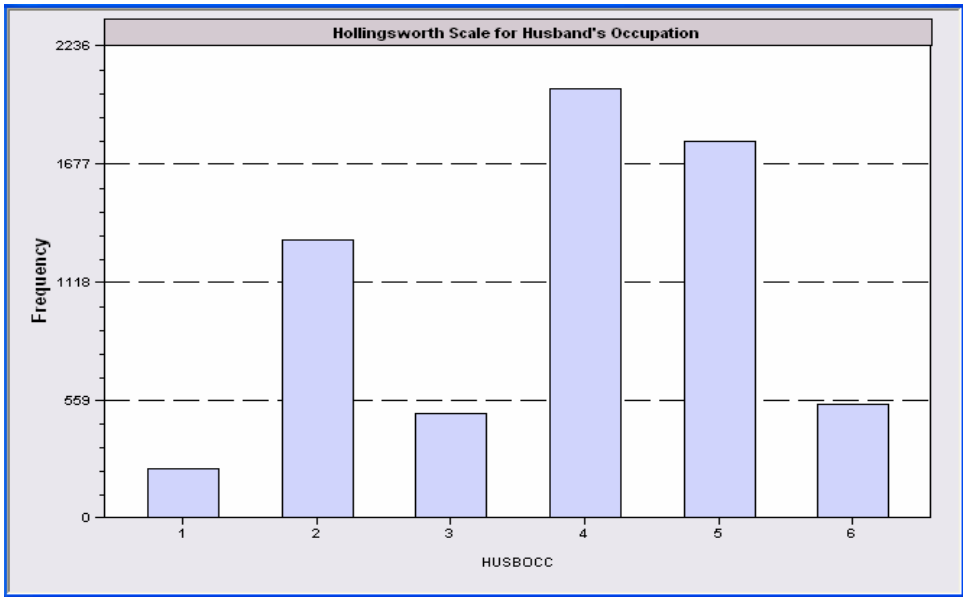


Figure 5.4a Sample proportions

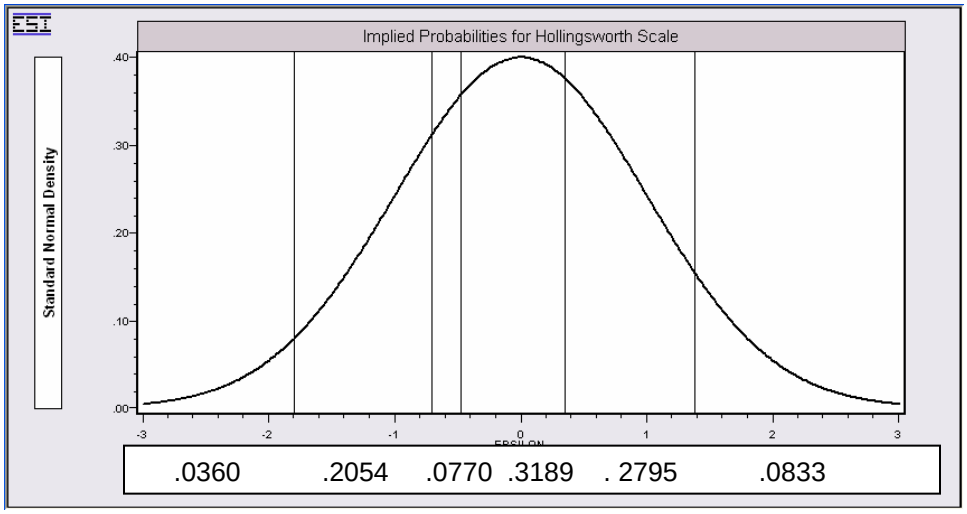


Figure 5.4b Implied Partitioning of Latent Normal Distribution

5.5 Interpretation of the Model – Partial Effects and Scaled Coefficients

Interpretation of the coefficients in the ordered probit model is more complicated than in the ordinary regression setting. [See, e.g., Daykin and Moffatt (2002).] There is no natural conditional mean function in the model. The outcome variable, y, is merely a label for the unordered, non-quantitative outcomes. As such, there is no conditional mean function, $E[y|\mathbf{x}]$ to analyze. (This is characteristic of discrete choice models.) In order to attach meaning to the parameters, one typically refers to the probabilities themselves. The partial effects in the ordered choice model are

$$\delta_j(\mathbf{x}_i) = \frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \left[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) \right]\boldsymbol{\beta}. \qquad \mathbf{P_2=.5091} \qquad 1)$$

A moment's inspection shows that neither the sign nor the magnitude of the coefficient is informative about the result above, so the direct interpretation of the coefficients is fundamentally ambiguous. [A counterpart result for a dummy variable in the model would be obtained by using a difference of probabilities, rather than a derivative. [See Boes and Winkelmann (2006a) and Greene (2007a, Chapter E22).] That is, suppose D is a dummy variable in the model (such as *Married*) and γ is the coefficient on D. We would measure the effect of a change in D from 0 to 1 with all other variables held at the values of interest (perhaps their means) using

$$\Delta_j(D) = \left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i + \gamma) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i + \gamma) \right] - \left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) \right].$$

(One might on occasion compute the partial effect for a dummy variable by differentiating as if it were a continuous variable. The results will typically resemble the finite change computation, sometimes surprisingly closely – the finite change is a discrete approximation to the derivative. Nonetheless, the latter computation is the more appropriate one.) The partial effects are shown in Table 5.2

The implication of the preceding result is that the effect of a change in one of the variables in the model depends on all the model parameters, the data, and which probability (cell) is of interest. It can be negative or positive. To illustrate, we consider a change in the education variable on the implied probabilities in Figure 5.3. Since the changes in a probability model are typically "marginal" (small), we will exaggerate the effect a bit so that it will show up in a figure. Consider, then, the same individual shown in Figure 5.3, except now, with a Ph.D. (college plus four years of postgraduate work). That is, 20 years of education, instead of the average 11.4 used earlier. The effect of an additional 8.6 years of education is shown in Figure 5.5. All five probabilities have changed. The two at the right end of the distribution have increased while the three at the left have decreased.

The partial effects give the impacts on the specific probabilities per unit change in the stimulus or regressor. For example, for continuous variable *Educ*, we find partial effects for the ordered probit model for the five cells of -.0034, -.00885, .00244, .00424, .00557, respectively, which give the expected change on the probabilities per additional year of education. For the income variable, for the highest cell, the estimated partial effect is .04055. However, some care is needed in interpreting this in terms of a unit change. The income variable has a mean of 0.34874 and a standard deviation of 0.1632. A full unit change in income would put the average individual nearly six standard deviations above the mean. Thus, for the marginal impact of income, one might want to measure a change in standard deviation units. Thus, an assessment of the impact of a change in income on the probability of the highest cell probability might be $0.04055 \times 0.1632 = 0.00662$. Precisely how this computation should be done will vary from one application to another.

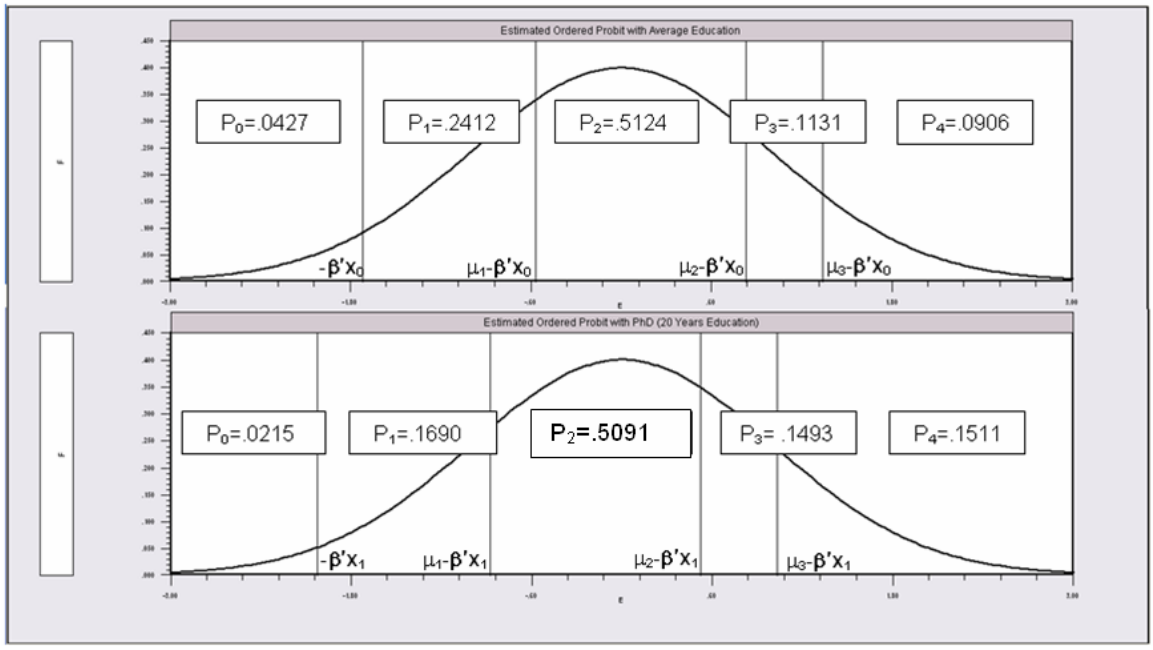


Figure 5.5 Partial Effect in Ordered Probit Model

Neither the signs nor the magnitudes of the coefficients are directly interpretable in the ordered choice model. It is necessary to compute partial effects of something similar to interpret the model meaningfully. In this computation, the only certainties in the signs of the partial effects in this model are as follows, where we consider a variable with a positive coefficient:

- Increases in that variable will increase the probability in the highest cell and decrease the probability in the lowest cell.
- The sum of all the changes will be zero. (The new probabilities must still sum to one.)
- The effects will begin at Pr(0) with one or more negative values, then change to a set of positive values; there will be one sign change. (This is the "single crossing" feature of the model. We will reconsider this aspect in Section 6.2.1.)

These are reversed for a variable with a negative coefficient.

One might also be interested in cumulative values of the partial effects, such as

$$\frac{\partial \text{Prob}(y \le j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \sum_{m=0}^{j} \left[f(\mu_{m-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_m - \boldsymbol{\beta}'\mathbf{x}_i) \right] \boldsymbol{\beta} . \tag{5.3}$$

See, e.g., Brewer et al. (2008). (Note that the last term in this set is zero by construction.) An example appears in Table 5.2.

Note in Table 5.1 there is a large difference in the coefficients obtained for the probit and logit models. The logit coefficients are roughly 1.8 times as large (not uniformly). This difference, which will always be observed, points up one of the risks in attempting to interpret directly the coefficients in the model. This difference reflects an inherent difference in the scaling of the underlying variable and in the shape of the distributions. The difference can be traced back (at least in part) to the different underlying variances in the two models. In the probit model, $\sigma_\varepsilon = 1$; in the logit model $\sigma_\varepsilon = \pi/\sqrt{3} = 1.81$. The models are roughly preserving the ratio $\boldsymbol{\beta}/\sigma_\varepsilon$ in the estimates. Note that the difference is greatly diminished (though not quite eliminated) in the partial effects reported in Table 5.2. That is the virtue of the scaling done to compute the

partial effects. The inherent characteristics of the model are essentially the same for the two functional forms.

Table 5.2 Estimated Partial Effects for Ordered Choice Models

Summary of Marginal Effects for Ordered Probability Model						
Effects computed at means. Effects for binary variables are						
computed as differences of probabilities, other variables at means.						
	Probit			Logit		
Outcome	Effect	dPy<=nn/dX	dPy>=nn/dX	Effect	dPy<=nn/dX	dPy>=nn/dX
	Continuous Variable AGE					
Y = 00	.00173	.00173	.00000	.00145	.00145	.00000
Y = 01	.00450	.00623	-.00173	.00521	.00666	-.00145
Y = 02	-.00124	.00499	-.00623	-.00166	.00500	-.00666
Y = 03	-.00216	.00283	-.00499	-.00250	.00250	-.00500
Y = 04	-.00283	.00000	-.00283	-.00250	.00000	-.00250
	Continuous Variable EDUC					
Y = 00	-.00340	-.00340	.00000	-.00291	-.00291	.00000
Y = 01	-.00885	-.01225	.00340	-.01046	-.01337	.00291
Y = 02	.00244	-.00982	.01225	.00333	-.01004	.01337
Y = 03	.00424	-.00557	.00982	.00502	-.00502	.01004
Y = 04	.00557	.00000	.00557	.00502	.00000	.00502
	Continuous Variable INCOME					
Y = 00	-.02476	-.02476	.00000	-.01922	-.01922	.00000
Y = 01	-.06438	-.08914	.02476	-.06908	-.08830	.01922
Y = 02	.01774	-.07141	.08914	.02197	-.06632	.08830
Y = 03	.03085	-.04055	.07141	.03315	-.03318	.06632
Y = 04	.04055	.00000	.04055	.03318	.00000	.03318
	Binary(0/1) Variable MARRIED					
Y = 00	.00293	.00293	.00000	.00287	.00287	.00000
Y = 01	.00771	.01064	-.00293	.01041	.01327	-.00287
Y = 02	-.00202	.00861	-.01064	-.00313	.01014	-.01327
Y = 03	-.00370	.00491	-.00861	-.00505	.00509	-.01014
Y = 04	-.00491	.00000	-.00491	-.00509	.00000	-.00509
	Binary(0/1) Variable KIDS					
Y = 00	-.00574	-.00574	.00000	-.00511	-.00511	.00000
Y = 01	-.01508	-.02081	.00574	-.01852	-.02363	.00511
Y = 02	.00397	-.01684	.02081	.00562	-.01801	.02363
Y = 03	.00724	-.00960	.01684	.00897	-.00904	.01801
Y = 04	.00960	.00000	.00960	.00904	.00000	.00904

5.5.1 Nonlinearities in the Variables

In the computation of partial effects, it is assumed that the independent variables can vary independently. When the model contains interactions of variables, or nonlinear functions of variables, the computation of partial effects becomes problematic, though more so in practice than in theory. [See Norton and Ai (2003) for extensive analysis of this issue.] Consider, for example, in our model if we added variables *EducSq* = *Educ*^{2} and *Educ***Age*. The estimated model is shown in Table 5.3 with some of the partial effects. Separate partial effects are shown for *Educ*, *Age*, *EducSq* and *EducAge*, as if they were independent variables. In fact, in this model, the partial effect for education would be

$$\delta_j(Educ) = \frac{\partial \text{Prob}(y = j \mid \mathbf{x})}{\partial Educ} = \left[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}) \right] \left(\beta_{Educ} + 2\beta_{EducSq} Educ + \beta_{EducAge} Age \right).$$

As Norton and Ai argued, none of the widely used computer packages computes this sort of result automatically. (It would be impossible for the software to anticipate every possible nonlinear function that might appear in the index function or recognize that function if it were implicit in a variable such as *EducAge*.) The analyst would have to compute this for themself. This can be computed using the results reported, as

$$\delta_j(Educ) = \frac{\partial \text{Prob}(y = j \mid \mathbf{x})}{\partial "Educ"} + (2Educ)\frac{\partial \text{Prob}(y = j \mid \mathbf{x})}{\partial "EducSq"} + Age\frac{\partial \text{Prob}(y = j \mid \mathbf{x})}{\partial "EducAge"}.$$

The derivatives shown for the zero cell in Table 5.3 are -.02028, .00057, .00004, respectively, and the means of *Age* and *Education* are 43.44 and 11.42, respectively. Thus, the partial effect for the probability of a zero outcome is -.00552. In our original model with the linear index function, the estimated effect was -.00340.

Table 5.3 Estimated Expanded Ordered Probit Model

	Expanded Ordered Probit				Ordered Probit				
	LogL = -5749.664				LogL = -5752.985				
	LogLR = -5752.985				LogL0 = -5875.096				
	Chisq = 6.642				Chisq = 244.2238				
	PseudoRqq = .0213499`				PseudoRsq= .0207847				
	Degrees of Freeedom 2				Degrees of Freedom 5				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.7422	.5520	1.344	.1788	1.9788	.1162	17.034	.0000	1.0000
AGE	-.0127	.0076	-1.664	.0961	-.0181	.0016	-11.166	.0000	43.4401
EDUC	.2124	.0709	2.995	.0027	.0356	.0071	4.986	.0000	11.4181
INCOME	.2583	.1044	2.474	.0134	.2587	.1039	2.490	.0128	.34874
MARRIED	-.0325	.0421	-.772	.4404	-.0310	.0420	-.737	.4608	.75217
KIDS	.0666	.0384	1.732	.0833	.0606	.0382	1.586	.1127	.37943
EDUCSQ	-.0060	.0023	-2.541	.0110					135.9773
EDUCAGE	-.0004	.0006	-.641	.5213					491.7343
Mu(1)	1.1495	.0212	54.288	.0000	1.1484	.0212	54.274	.0000	
Mu(2)	2.5501	.0216	117.914	.0000	2.5478	.0216	117.856	.0000	
Mu(3)	3.0589	.0265	115.561	.0000	3.0564	.0267	115.500	.0000	

Marginal Effects for Ordered Probit Model

Outcome	AGE	EDUC	EDUCSQ	EDUCAGE
Y = 00	.00121	-.02028	.00057	.00004
Y = 01	.00316	-.05290	.00148	.00010
Y = 02	-.00087	.01458	-.00041	-.00003
Y = 03	-.00151	.02534	-.00071	-.00005
Y = 04	-.00198	.03326	-.00093	-.00007

5.5.2 Average Partial Effects

In computing partial effects, we have evaluated the functions by inserting the sample means of the regressors. That is, our computation for *Educ*, for example, is

$$\frac{\partial \text{Prob}(y = j \mid \bar{\mathbf{x}})}{\partial Educ} = \left[f(\mu_{j-1} - \boldsymbol{\beta}'\bar{\mathbf{x}}) - f(\mu_j - \boldsymbol{\beta}'\bar{\mathbf{x}}) \right] \beta_{Educ}.$$

The *average partial effect*, or APE, is computed instead by evaluating the partial effect for each individual and averaging the computed effects. thus,

$$APE_j(Educ) = \frac{1}{n} \sum_{i=1}^{n} \left[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) \right] \beta_{Educ}.$$

In practice, unless the sample size is very small or the data are highly skewed and affected by outliers, this will give a very similar result. For the example suggested, the first computation gives ∂Prob(y=4)/∂Educ = 0.005557 (see Table 5.2 for the probit model) and the second gives 0.005723, a difference of about 2.7%. Further discussion of the computation of APEs and standard errors using the delta method appear in Greene (2008a, pp. 783-785).

5.5.3 Interpreting the Threshold Parameters

In most treatments, the threshold parameters, μ_j are treated as nuisance parameters; necessary for the computations, but of no intrinsic interest on their own. Daykin and Moffatt (2002,p. 162) argue that in psychology applications with attitude scales, "If the statement is one with which most people are either in strong agreement or strong disagreement, we would expect the cut points to be tightly bunched in the middle of the distribution. If, in contrast, the statement is one on which people are not keen to be seen expressing strong views, we would expect the cut points to be more widely dispersed." Thus, in the absence of other information, this suggests that the threshold parameters can reveal some information about the preferences of the respondents. [In contradiction, Anderson (1984, p. 4) states "The estimates of the θ_s are strongly related to the average proportion in the corresponding categories, as recourse to any specified functional form for $F(.)$ indicates. (See the example in Section 5.4.) Hence, the θ_s parameters are not informative about the closeness of categories. As noted above, the regression relationship is based on $\boldsymbol{\beta}'\mathbf{x}$ and is firmly one dimensional."

5.5.4 The Underlying Regression

One would typically not be interested in the underlying regression. The observed variable will always be the discrete, ordered outcome. Nonetheless, the model does imply a set of partial changes for the latent regressand,

$$\partial E[y^* \mid \mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}.$$

This differs from more familiar cases in that the scaling of the dependent variable has been lost due to the censoring. Thus, it is impossible to attach any meaning to the change in the mean. McElvey and Zavoina (1975) suggest that if one is going to base interpretation of the model on the latent regression, then the coefficients should be "standardized." That is, changes should be measured in standard deviation units. A standardized regression coefficient for variable *k* would

be

$$\beta_k^* = \beta[s_{kk} / s_{y^*}],$$

where s_{kk} is the standard deviation of the regressor of interest and s_{y^*} is the standard deviation of y^*. Measurement of s_{kk} is straightforward based on the observed data. For s_{y^*}, the authors suggest the computation be based on the implication of the regression;

$$y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$$

so

$$\text{Var}[y^*] = \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{xx}} \boldsymbol{\beta} + \sigma_\varepsilon^2. \tag{5.4}$$

The two components are easily computed using the observed data and the normalized value of σ_ε^2, 1.0 or $\pi^2/3$. For our ordered logit model in Table 5.1, the estimate of s_{y^*} is 1.03156. The results of the computation are shown in Table 5.4

Table 5.4 Transformed Latent Regression Coefficients

Variable	β	β*
Age	-.01808	-2.23279
Educ	.03556	.19325
Income	.25869	.00676
Married	-.03100	-.00560
Kids	.06065	.01385

Some caution is needed when interpreting these. The variable that is assumed to be changing is an underlying preference scale. The notion of a unit or standard deviation change in utility or feeling is a bit dubious. That is among the motivations for discrete choice analysis of this sort; it frees the analyst from having to attach units of measure to unmeasurable quantities while still enabling them to learn about important features of preferences.

5.6 Inference

This section considers hypothesis tests about model components.

5.6.1 Inference about Coefficients

The model has been fit by maximum likelihood. The estimates are shown in Table 5.1. The assumptions underlying the regularity conditions for maximum likelihood estimation should be met, so inference can be based on conventional methods. Standard errors for the estimated coefficients are computed by inverting an estimator of the negative of the expected second derivatives of the log likelihood. This will either be based on the actual second derivatives,

$$\begin{aligned}\mathbf{V}_H = Est.Asy.Var\begin{bmatrix}\hat{\boldsymbol{\beta}}_{MLE}\\ \hat{\boldsymbol{\mu}}_{MLE}\end{bmatrix} &= \left[-\sum_{i=1}^{N}\frac{\partial^2 \log \Pr(y=y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\mu}}_{MLE})}{\partial\begin{bmatrix}\hat{\boldsymbol{\beta}}_{MLE}\\ \hat{\boldsymbol{\mu}}_{MLE}\end{bmatrix}\partial\begin{bmatrix}\hat{\boldsymbol{\beta}}'_{MLE} & \hat{\boldsymbol{\mu}}'_{MLE}\end{bmatrix}}\right]^{-1} \\ &= \left[-\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right]^{-1} \\ &= \left[-\hat{\mathbf{H}}\right]^{-1},\end{aligned} \tag{5.5}$$

or the sum of the outer products of the first derivatives (the BHHH or outer product of gradients, OPG, estimator),

$$\begin{aligned}\mathbf{V}_{OPG} &= \left[\sum_{i=1}^{N}\left(\frac{\partial \log \Pr(y=y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\mu}}_{MLE})}{\partial \begin{bmatrix}\hat{\boldsymbol{\beta}}_{MLE}\\ \hat{\boldsymbol{\mu}}_{MLE}\end{bmatrix}}\right)\left(\frac{\partial \log \Pr(y=y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\mu}}_{MLE})}{\partial \begin{bmatrix}\hat{\boldsymbol{\beta}}_{MLE}\\ \hat{\boldsymbol{\mu}}_{MLE}\end{bmatrix}}\right)'\right]^{-1} \\ &= \left[\sum_{i=1}^{N}\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i'\right]^{-1} \\ &= \left[\hat{\mathbf{G}}'\hat{\mathbf{G}}\right]^{-1}.\end{aligned} \tag{5.6}$$

Generally, two procedures, the Wald test and the likelihood ratio test are used for testing hypotheses. A third, the LM test, is available, but rarely used because of its complexity compared to the other two.

Inference about a single coefficient is based on the standard "*z*" test. The test of a simple null hypothesis:

$$H_0: \beta_k = \beta_k^0,$$

is tested by referring the Wald statistic,

$$z = \frac{\hat{\beta}_{k,MLE} - \beta_k^0}{Est.Std.Err(\hat{\beta}_{k,MLE})},$$

to a table of the standard normal distribution. Estimated standard errors are obtained as the square roots of the diagonals of the matrix described in the previous paragraph. For example, the conventional test against the null hypothesis H_0:$\beta_k = 0$ is reported as standard results when the model is estimated. The test is carried out in the results shown in Table 5.1 for the estimated model, where we find that *Age*, *Educ* and *Income* are "significant" determinants of the probabilities while *Married* and *Kids* are not.

Inference about the threshold parameters would be meaningless, and is not generally carried out. In the results below, we find a typical pattern; the threshold parameters have very small standard errors and are "highly significant." Note, however, that a test of the hypothesis that $\mu_2 = 0$, would not be useful because μ_2 must be greater than μ_1 and μ_0, and $\mu_0 = 0$. Without this ordering, the model becomes internally inconsistent – the probabilities can be negative.

A test about more than one coefficient can be carried out using a Wald test. For a null hypothesis of the form

$$H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{q},$$

where **R** is a matrix of coefficients in the linear restrictions and **q** is a vector of constants, the statistic will be

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}\right)'\left[\mathbf{R}\mathbf{V}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}\right),$$

where **V** is the estimated asymptotic covariance matrix of the coefficients. The difficulty of this computation will vary from one program to another. Both *Stata* and *NLOGIT* have built in "*Wald*" commands that can be used to do the computation as well as matrix algebra routines that also allow the user to program the computation themselves. For example, the following tests the null hypothesis that the coefficients on *EducSq* and *EducAge* in our expanded model in Table 5.3 are simultaneously zero. As noted, the statistic is treated as a chi squared statistic with degrees of freedom equal to the number of restrictions. In the results below, for example, we see that we would reject the hypothesis that both are zero, evidently because of the significance of the first one.

```
Ordered ; Lhs = Health
        ; Rhs = one,age,educ,income,married,kids,educsq,educage $
Wald    ; fn1 = b_educsq ; fn2 = b_educag $
+---------------------------------------------+
| WALD procedure. Estimates and standard errors |
| for nonlinear functions and joint test of     |
| nonlinear restrictions.                       |
| Wald Statistic             =      6.64372     |
| Prob. from Chi-squared[ 2] =       .03609     |
+---------------------------------------------+
+--------+--------------+----------------+--------+--------+
|Variable| Coefficient  | Standard Error |b/St.Er.|P[|Z|>z]|
+--------+--------------+----------------+--------+--------+
|Fncn(1) |    -.00596**        .00234587    -2.541   .0110 |
|Fncn(2) |    -.00042          .00065479     -.641   .5213 |
+--------+-----------------------------------------------+
```

The counterparts for this computation in *Stata* would be

```
. oprobit health age educ income married kids educsq educage
. test educsq educage
( 1) [health]educsq = 0
( 2) [health]educage = 0
           chi2(  2) =   6.644
         Prob > chi2 =   0.0361
```

The computation can be programmed directly using matrix algebra, e.g., with *NLOGIT* as

```
Matrix        ; b2=b(7:8);v22=varb(7:8,7:8) $
Matrix        ; list ; Wald = b2'<v22>b2 $
Matrix WALD     has  1 rows and  1 columns.
               1
        +--------------
       1|    6.64372
```

or using the *Mata* package in *Stata* or *PROC MATRIX* in *SAS*. In any case, using the built in procedure has the advantage of producing the "*p*-value" for the statistic as well as the statistic itself.

The likelihood ratio test will usually be simpler than the Wald test if the hypothesis is more involved than the simple zero restrictions shown above, though it does require estimation of both the null (restricted) and alternative (unrestricted) models. The test statistic is simply twice the difference between the log likelihoods for the null and alternative models. For the earlier example, the log likelihood for the (alternative) model that includes *EducSq* and *EducAge* is -5749.664 while, as seen earlier, the log likelihood for the (null) model that omits these variables is -5752.985. The test statistic is

$$LR = 2(-5749.664 - (-5752.985)) = 6.642.$$

This is nearly the same as the Wald statistic and produces the same conclusion. The two tests could conflict for a particular significance level. This is a finite sample result – asymptotically, the two statistics have the same characteristics when the assumptions of the model are met. As a general occurrence (albeit not necessarily), the Wald statistic will usually be larger than the *LR* statistic. Purely heuristically, because it uses more information – it is based on both models – we prefer the *LR* statistic.

A common test of the sort considered here is a "test of the model" in the spirit of the overall *F* statistic in the linear regression model that is used to test the null hypothesis that all coefficients in the model save the constant term are zero. The counterpart for the ordered choice model would be likelihood ratio test against the null hypothesis that the model contains only a constant term and the threshold parameters. This test statistic is routinely reported with the standard results for the estimated model by all commercial packages. For the results in Table 5.3, we have a model chi squared of 244.2238 with five degrees of freedom.

Note it is not necessary to estimate the null model to carry out this test. The maximum likelihood estimates of the parameters of the model when it contains only a constant term are equivalent to method of moments estimators based on the following moment equations involving the raw sample proportions:

$P_0 = \Pr(y = 0) = F(-\alpha)$
$P_1 = \Pr(y \leq 1) = F(\mu_1 - \alpha)$
$P_j = \Pr(y \leq j) = F(\mu_j - \alpha)$
and so on.

These can be solved directly, in the logit case using a hand calculator (e.g., $a = \log(P_0/(1-P_0))$. These (with $\beta = 0$) are the usual starting values for the iterations, so the log likelihood computed at entry to the iterative procedure provides the needed value for the null model.

5.6.2 Testing for Structural Change or Homogeneity of Strata

The likelihood ratio test provides a more convenient approach for testing homogeneity of strata in the data. For example, our data are separated by men and women in the introduction, and one might be interested in testing whether the same model should be used to describe the two groups. The counterpart to a "Chow test" [Chow (1960), Greene (2008, p. 121)] in linear regression would be a test of group homogeneity in the choice model. The test statistic is easily computed using

$$LR = 2[\Sigma_{g=groups} \log L_g - \log L_{pooled}].$$

The statistic has a limiting chi squared distribution with degrees of freedom equal to *G*-1 times the number of parameters in the model (slopes and thresholds). Our data are segmented by gender in the introduction. For a test of the null hypothesis that the same ordered choice model applies to the two groups, we find $\log L_{Male} = -2952.05$, $\log L_{Female} = -2798.03$ and $\log L_{Pooled} = -5752.985$. Applying the preceding result gives a chi squared value of 5.83 with 9 degrees of freedom. The *p*-value is 0.7569 (the 95% critical value is 16.92). On this basis we conclude that is appropriate to pool these two subsamples. (In RWM's analysis, they maintained the sample division. However, they were not analyzing the health satisfaction variable.)

5.6.3 Robust Covariance Matrix Estimation

There are two candidates available for the estimated asymptotic covariance matrix of the parameter estimators, $-\mathbf{H}^{-1}$ based on the Hessian and $(\mathbf{G}'\mathbf{G})^{-1}$ based on the first derivatives. [See Section 5.6.1, (5.5) and (5.6).] The implication of the *Information Matrix Equality* [see Greene (2008a, Ch. 16)] is that these two matrices estimate the same covariance matrix and are, for practical purposes, interchangeable. A third matrix, the "robust" covariance matrix is often computed in recent applications, that being

$$\mathbf{V}_R = [-\mathbf{H}^{-1}] (\mathbf{G}'\mathbf{G}) [-\mathbf{H}^{-1}]. \tag{5.7}$$

The logic of the computation can be seen by assuming that Netwon's method is used to estimate the parameters. The maximum likelihood estimator at the maximum will produce

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}^0\right) = \left[-\hat{\mathbf{H}}/n\right]^{-1}\left(\sqrt{n}\sum_{i=1}^{n}\mathbf{g}_i\right) + o(1/n),$$

where $\boldsymbol{\theta}^0$ is the vector of parameters that the MLE converges to and $o(1/n)$ denotes a trailing term that converges to zero as $n \to \infty$. The asymptotic variance of the MLE is obtained by multiplying the limiting variance of the right hand side by $1/n$. The trailing terms will disappear. The leading matrix in brackets converges (we assume) to its expectation – a constant matrix. For the vector in parentheses, if the model assumptions are correct, then by the information matrix equality, its limiting variance will be $-\mathbf{H}/n$. Two occurrences of $\mathbf{H}$ will cancel and we are left with $\mathbf{V}_H$ as the usual estimator. But, ignoring the information matrix equality, whether it is met or not, the asymptotic variance of the MLE will be estimable by using $(1/n)\mathbf{G}'\mathbf{G}$ as an estimator of the variance matrix of the quantity in parentheses. Then, the "robust" covariance matrix estimator becomes the *sandwich estimator* given above.

This produces two cases: If the model assumptions are correct, then the robust estimator is the same as either of the conventional estimators. If the model assumptions are incorrect, then the robust estimator still produces the asymptotic covariance matrix for the MLE. (A familiar application of this result is the "White" (1980) estimator for the asymptotic covariance matrix of the OLS estimator in the presence of heteroscedasticity.) But, a new question arises in the second case. If the model assumptions are not correct, then what is $\boldsymbol{\theta}^0$? In order for this computation to be useful, it must be the case that in spite of the failure of the model assumptions, $\hat{\boldsymbol{\theta}}_{MLE}$ must still be a consistent estimator of the parameters of interest, in the present case, $(\boldsymbol{\beta}', \boldsymbol{\mu}')'$. Once again, the case of OLS in the presence of heteroscedasticity provides a useful benchmark. On the other hand, for the ordered probit model, any of the following will render the estimator of the parameters inconsistent: (i) omitted variables even if they are orthogonal to included variables, (ii) heteroscedasticity in ε, (iii) incorrect distributional assumption – e.g., using the logit model when the probit model is the correct one, (iv) endogeneity of any of the regressors, (v) omission of latent heterogeneity – this is equivalent to an omitted variable. Indeed, it is difficult to produce a model failure that the estimator is robust to. One possibility that seems unlikely in this cross section setting is correlation across observations. The upshot is that either the "robust covariance matrix" estimator is the same as the other two already considered, or it is a "robust" covariance matrix for an inconsistent estimator of the parameters. [Additional commentary on this result appears in Freedman (2006).]

5.6.4 Inference About Partial Effects

Partial effects are computed using either the derivatives or first differences for discrete variables;

$$\boldsymbol{\delta}_j(\mathbf{x}_i) = \frac{\partial \text{Prob}(y=j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \left[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) \right] \boldsymbol{\beta}, \tag{5.8}$$

$$\Delta_j(d,\mathbf{x}_i) = \left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i + \gamma) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i + \gamma) \right] - \left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) \right].$$

Since these are functions of the estimated parameters, they are subject to sampling variability and one might desire to obtain appropriate asymptotic covariance matrices and/or confidence intervals. For this purpose, the partial effects are typically computed at the sample means. [See Greene (2008a, pp. 780-785) for analysis of this computation for average partial effects.] The delta method is used to obtain the standard errors. Let $\mathbf{V}$ denote the estimated asymptotic covariance matrix for the $(K+J-2)\times 1$ parameter vector $\left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\mu}}'\right)'$. Then, the estimator of the asymptotic covariance matrix for each vector of partial effects is

$$\mathbf{Q} = \hat{\mathbf{C}}\mathbf{V}\hat{\mathbf{C}}',$$

where

$$\hat{\mathbf{C}} = \left[\frac{\partial \hat{\boldsymbol{\delta}}_j(\bar{\mathbf{x}})}{\partial \hat{\boldsymbol{\beta}}'} \quad \frac{\partial \hat{\boldsymbol{\delta}}_j(\bar{\mathbf{x}})}{\partial \hat{\boldsymbol{\mu}}'} \right].$$

The appropriate row of $\hat{\mathbf{C}}$ is replaced with the derivatives of $\Delta_j(d, \bar{\mathbf{x}})$ when the effect is being computed for a discrete variable.

Patterns of statistical significance for the partial effects will usually echo those for the coefficients themselves. This will follow from the fact that $\mathbf{C}$ is of the form

$$\mathbf{C} = [a_{ij}\mathbf{I}, \mathbf{0}] + [\mathbf{C}_{\boldsymbol{\beta}}, \mathbf{C}_{\boldsymbol{\mu}}],$$

where a_{ij} is the bracketed scalar term in $\hat{\boldsymbol{\delta}}_j(\bar{\mathbf{x}})$. The second matrix is typically much smaller than the first. Thus, the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\delta}}_j(\bar{\mathbf{x}}) = a_{ij}\boldsymbol{\beta}$ typically resembles $a_{ij}^2\mathbf{V}$. The scale factor would cancel out of a "z value" leaving the typical result. It is clearly visible in the results in Table 5.5. This result does raise a vexing question. It is conceivable for the significance tests of $\delta_j(x_k)$ to conflict with each other, that is, with $\delta_m(x_k)$ for an $m \neq j$, and/or with a test about the associated coefficient, β_k. Since $\delta_j(x_k) = a_{ij}\beta_k$, the tests would seem to be in direct contradiction. The natural question for the practitioner, then, is where should the appropriate test of significance be carried out. Opinions differ and there is no single answer. It might logically be argued that the overall purpose of the regression analysis is to compute the partial effects, so that is where the tests should be carried out. On the other hand, the meaning of the test with respect to the partial effects is ambiguous, since they are functions of all the parameters as well as the data. The number of possible contradictions is large. Our preference on the methodological basis is for the structural coefficients, not the partial effects.

Table 5.5 Estimated Partial Effects with Asymptotic Standard Errors

Marginal effects for ordered probability model				
M.E.s for dummy variables are Pr[y|x=1]-Pr[y|x=0]				
Names for dummy variables are marked by *.				
Variable	Coefficient	Standard Error	b/St.Error	P[|Z|>z]
	These are the effects on Prob[Y=00] at means.			
AGE	.00173	.000165	10.488	.0000
EDUC	-.00340	.000692	-4.919	.0000
INCOME	-.02476	.009973	-2.483	.0130
*MARRIED	.00293	.003920	.747	.4551
*KIDS	-.00574	.003578	-1.603	.1089
	These are the effects on Prob[Y=01] at means.			
AGE	.00450	.000403	11.161	.0000
EDUC	-.00885	.001775	-4.986	.0000
INCOME	-.06438	.025851	-2.490	.0128
*MARRIED	.00771	.010440	.738	.4604
*KIDS	-.01508	.009494	-1.588	.1122
	These are the effects on Prob[Y=02] at means.			
AGE	-.00124	.000170	-7.310	.0000
EDUC	.00244	.000549	4.438	.0000
INCOME	.01774	.007356	2.411	.0159
*MARRIED	-.00202	.002611	-.774	.4387
*KIDS	.00397	.002419	1.641	.1009
	These are the effects on Prob[Y=03] at means.			
AGE	-.00216	.000241	-8.958	.0000
EDUC	.00424	.000901	4.709	.0000
INCOME	.03085	.012559	2.457	.0140
*MARRIED	-.00370	.005033	-.736	.4620
*KIDS	.00724	.004599	1.574	.1154
	These are the effects on Prob[Y=04] at means.			
AGE	-.00283	.000271	-10.452	.0000
EDUC	.00557	.001130	4.931	.0000
INCOME	.04055	.016335	2.482	.0130
*MARRIED	-.00491	.006733	-.729	.4657
*KIDS	.00960	.006120	1.569	.1166

5.7 Prediction – Computing Probabilities

One might want to use the model for prediction as well as inference. The natural predictor would seem to be $\hat{y}^* = \hat{\boldsymbol{\beta}}'\mathbf{x}$. However, the underlying variable is typically unobservable, and often of no intrinsic interest in its own right. (E.g., in the bioassay case, the "tolerance" of a particular insect would probably be of little interest. In the preference scale case such as in our health satisfaction example, the underlying utility is inherently unmeasurable.) The more natural exercise would be to predict the observed outcome. Since it is discrete, the linear predictor is of little use. The starting point would be the predicted probabilities. The model provides predictors

$$\begin{aligned}\hat{P}_j(\mathbf{x}_i) &= F\left(\hat{\mu}_j - \hat{\boldsymbol{\beta}}'\mathbf{x}_i\right) - F\left(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}'\mathbf{x}_i\right) \\ &= \hat{F}_{j,i} - \hat{F}_{j-1,i}, j = 0,1,...,J.\end{aligned} \tag{5.9}$$

If the sample is small enough and particular observations are of interest, a simple listing might be useful. For our sample of 4,483 observations, this would probably not be helpful. One might, instead, tabulate predicted probabilities against variables of interest. For example, for reasons unknown to us, the presence of children in the household appears to have a substantial

(increasing) impact on whether one reports the lowest value of health satisfaction. A set of results is shown in Table 5.6.

Table 5.6 Mean Predicted Probabilities by Kids

Variable	Mean	Std.Dev.	Minimum	Maximum
Stratum is KIDS = 0.000. Nobs.= 2782.000				
P0	.059586	.028182	.009561	.125545
P1	.268398	.063415	.106526	.374712
P2	.489603	.024370	.419003	.515906
P3	.101163	.030157	.052589	.181065
P4	.081250	.041250	.028152	.237842
Stratum is KIDS = 1.000. Nobs.= 1701.000				
P0	.036392	.013926	.010954	.105794
P1	.217619	.039662	.115439	.354036
P2	.509830	.009048	.443130	.515906
P3	.125049	.019454	.061673	.176725
P4	.111111	.030413	.035368	.222307
All 4483 observations in current sample				
P0	.050786	.026325	.009561	.125545
P1	.249130	.060821	.106526	.374712
P2	.497278	.022269	.419003	.515906
P3	.110226	.029021	.052589	.181065
P4	.092580	.040207	.028152	.237842

Standard errors and confidence intervals can be computed using the delta method. These are a bit simpler than for the partial effects, as there is no need to make a distinction between discrete and continuous variables. The matrix of derivatives has a row for each outcome, containing

$$\frac{\partial \hat{P}_j(\mathbf{x}_i)}{\partial(\hat{\boldsymbol{\beta}}' \quad \hat{\boldsymbol{\mu}}')} = \left[\left(\hat{f}_{j-1}(\mathbf{x}_i) - \hat{f}_j(\mathbf{x}_i)\right)\mathbf{x}_i' \quad \left(0,...,-\hat{f}_{j-1},\hat{f}_j,0,...\right)\right]. \tag{5.10}$$

For certain variables of interest, a plot of the predicted probabilities against the values of the variable might be useful. In our application, *Age* seems to be an important determinant of self assessed health satisfaction. A plot of the predicted probabilities for this model for the values of *Age* in the sample, 25 to 64, for a person who has average income and education, and is married with children appears in Figure 5.6.

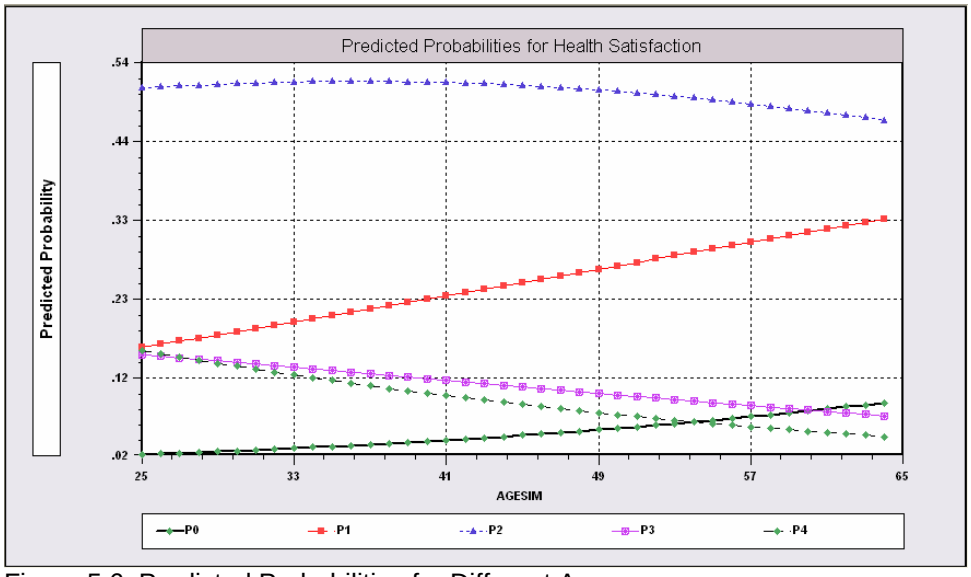


Figure 5.6 Predicted Probabilities for Different Ages

5.8 Measuring Fit

The search for a scalar measure of model fit for discrete choice models must be among the least satisfying of the exercises in the modeling effort. Superficially, the search is for a counterpart to the R^2 = "proportion of the variation in the dependent variable that is explained by variation in the independent variables." The search is frustrated in this (and other discrete choice models) for two reasons:

- There is no "dependent variable." In the ordered choice model, there are J+1 explained variables that are defined by $m_{ij} = 1$ if $y_i = j$ and 0 otherwise and which satisfy the constraints $m_{ij} = 0$ or 1 and $\Sigma_j m_{ij} = 1$. (This is true for the bioassay case as well; the observed proportions for each i consist of the sample means of m_{ij} for n_i observations with a common $\mathbf{x}_i$.) The observed variable y_i is nothing more than a labeling convention for the regions of the real line defined by the partitioning in the model specification.

- There is no "variation" (around the mean) to be explained. The outcome is not a measure of a quantity; it is a label. There is no conditional mean, as such, either.

For these reasons, one needs to exert a considerable amount of caution in computing and reporting "measures of fit" in this setting.

A "fit measure" that one computes can be used for two purposes: (i) to assess the fit of the predictions by the model to the observed data, compared to no model and (ii) to compare the model one estimates to a different model. For the first of these, we (and a generation of others) have suggested the overall model chi squared,

$$\chi^2[K+J-2] = 2[\log L_{Model} - \log L_{No\ Model}].$$

A transformation of this statistic that is (very) often reported in the contemporary literature is McFadden's (1977) "pseudo R^2" which is computed as

$$R_{Pseudo}^2 = 1 - \log L_{Model} / \log L_{No\ Model}.$$

A degrees of freedom adjusted version is sometimes reported,

$$Adjusted\ R_{Pseudo}^2 = 1 - [\log L_{No\ Model} - M] / \log L_{Model}.$$

where M is the number of parameters in the model. The pseudo R^2 has the virtues that it is bounded by 0 and 1, and increases whenever the model increases in size – that is, the pseudo R^2 is larger for any model compared to a model that is nested within it. It is important to emphasize, as is clear from the definition, it is not a measure of model fit to the data and it is not a measure of the proportion of variation explained in any sense. (It is also worth noting that it is not necessarily bounded by zero and one unless the model in question is a discrete choice model for which the log likelihood function is necessarily negative. For example, it is a simple exercise to show that the log likelihood for a linear normal regression model can be positive or negative, depending on the value of σ_ε, which could produce values outside the unit interval.) Lastly, the *Pseudo* R^2 cannot reach one, though it can equal zero.

The value of the *Pseudo* R^2 in the model we have analyzed above can be found in Table 5.1 for the basic model (0.0207847) and in Table 5.4 for the expanded model (0.02135). The low values might seem a bit surprising given the several highly significant coefficient estimates in the reported results. However, as with the counterpart in linear regression, highly significant coefficients need not attend a high fit measure.

A second measure for the ordered choice model was suggested by McKelvey and Zavoina (1975). The logic of their measure is based on predicting the underlying latent variable, y^*. The total variance *in the underlying variable* in the ordered choice model is

$$\text{Var}[y^*] = \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{xx}} \boldsymbol{\beta} + \sigma_\varepsilon^2.$$

where $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is the theoretical covariance matrix of $\mathbf{x}_i$. The first part of this is estimable using the maximum likelihood estimates of $\boldsymbol{\beta}$ and the sample covariance matrix for the data, and the second part is known to be 1.0 or $\pi^2/3$ for the probit and logit models, respectively. Thus, the authors suggested

$$R_{MZ}^2 = 1 - \frac{\sigma_\varepsilon^2}{\hat{\boldsymbol{\beta}}' \mathbf{S}_{\mathbf{xx}} \hat{\boldsymbol{\beta}} + \sigma_\varepsilon^2}.$$

They defined the "explained" part of this computation in terms of deviations from a prediction, $e_i = \hat{y}_i - \hat{\bar{y}}$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}' \mathbf{x}_i$, producing

$$R_{MZ}^2 = \frac{\Sigma_{i=1}^n (\hat{y}_i - \hat{\bar{y}})^2}{\Sigma_{i=1}^n (\hat{y}_i - \hat{\bar{y}})^2 + n}.$$

With this computation, we obtain an improvement over the *PseudoR*^{2}; for our model, R_{MZ}^2 = 0.06024.

Long and Freese (2006) list a variety of other measures that are computed for the ordered choice models. (This set of results is produced by a *Stata* program called `FitStat` written by one of the authors. We mention it at this juncture to illustrate the problem of searching for a fit measure in a particular discrete choice model, not to recommend that analysts either do or do not use it or these results. The formulas below do not appear in Long and Freese or in the

documentation for *Stata*; they are described in long detail by UCLA/ATS (2008) among others and, of course, piecemeal by the original designers.) These include

$$R^2_{Cox,Snell} = 1 - \left[\frac{\log L_{No\ Model}}{\log L_{Model}} \right]^{2/n},$$

$$R^2_{Cragg,Uhler/Nagelkerke} = \frac{1 - \left[\frac{\log L_{No\ Model}}{\log L_{Model}} \right]^{2/n}}{1 - [\log L_{No\ Model}]^{2/n}}$$

In UCLA/ATS (2008), it is noted that "pseudo *R*-squareds" for categorical variables serve three functions:

> Measures of explained variability,
> Measures of improvement from null model to fitted model,
> Square of the correlation.

None of the already suggested fit measures bear any relation to the first and third of these. All are connected to the improvement in the log likelihood by the addition of the variables in the model to a constants only model. Of course, the log likelihood functions, themselves, do that, and what these statistics add to the two values is a transformation that is between zero and one. It is worth noting, the measures are strictly between zero and one. None can achieve one even if the model predicts perfectly (somehow – we have not defined what would be meant by "predict"). Nonetheless, what they do all share is that they increase as the model grows and they are bounded by zero and one. (However, the "adjusted pseudo R^2" can decline as variables are added, in the same fashion as $\bar{R}^2$ for linear regression.)

UCLA/ATS (2008) observe (with reference to a binary logit model),

> When analyzing data with a logistic regression, an *equivalent statistic to R-squared does not exist*. [Emphasis added.] The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo *R*-squareds have been developed. These are "pseudo" *R*-squareds because they look like *R*-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo *R*-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS *R*-squared and different pseudo R-squareds can arrive at very different values

The notion of "model fit" in this and elsewhere relates to the log likelihood for the model, not to an assessment of how well the model predicts the outcome variable, as it does in regression analysis.

It seems appropriate to add a fourth item to the list above; fit measures are used to compare models to each other, not only to baseline, "null" models. For this purpose, a handful of other fit measures that are not normalized to the unit interval, but are based on the log likelihood function, are often used:

$$\textit{Log Akaike Information Criterion} = AIC = [-2\log L + 2M]/n, \quad (5.11)$$

$$\textit{Finite Sample AIC} = AIC_{FS} = AIC + 2M(M+1)/(n - M - 1),$$

$$\textit{Bayes Information Criterion} = BIC = [-2\log L + M/\log n]/n,$$

$$\textit{Hannan-Quinn IC} = HQIC = [-2\log L + 2\, M \log\log n]/n.$$

The information measures are all created in the spirit of adjusted R^2 – they reward a model for "fit" with few parameters and small samples. A better model is one with a smaller information criterion. (Long and Freese mention two others, "*AIC used by Stata*" and "*BIC used by Stata.*" We have been unable to decipher what these are.)

Long and Freese (p. 196) and UCAL/ATS (2008) mention two other measures that seem (to these authors) to have received far less attention than these likelihood based measures. These are

$$\textit{Count } R^2 = \frac{\textit{Number of Correct Predictions}}{n}$$

and

$$\textit{Adjusted Count } R^2 = \frac{\textit{Number of Correct Predictions} - n_j^*}{n - n_j^*}.$$

Where n_j* is the count of the most frequent outcome. The discussion is about binary choice models, so we have to extend the idea to our ordered choice model. There is a long catalog of fit measures for binary choice models based on this sort of computation. [See, e.g., Greene (2008a, pp. 790-793).] The central feature is a fitting mechanism: Predict $y = j$ if the model states that j is the most likely outcome. In the binary choice case, the rule is to use as the prediction, the outcome which has probability exceeding 0.5. For the ordered choice case, this would suggest using the rule

$$\hat{y}_i = j^* \textit{ such that estimated } \text{Prob}(y_i = j^* \mid \mathbf{x}_i) > \Pr(y_i = j \mid \mathbf{x}_i)\ \forall\ j \neq j^*.$$

That is, put the predicted y in the cell with the highest probability. This rule has an aesthetic appeal, and in the absence of priors (as in a Bayesian setting) we have not found a preferable approach. Nonetheless, this can lead to an unexpected outcome. For our first example in Table 5.1, this rule produces the results in Table 5.7.

Table 5.7 Predicted vs. Actual Outcomes for Ordered Probit Model

Cross tabulation of predictions.						
Row is actual, column is predicted.						
Model=Probit. Prediction=most likely cell.						
Actual	0	1	2	3	4	Row Sum
0	0	0	230	0	0	220
1	0	0	1113	0	0	1113
2	0	0	2226	0	0	2226
3	0	0	500	0	0	500
4	0	0	414	0	0	414
Col Sum	0	0	4483	0	0	4483

By this method, our model, with its highly significant overall fit and several highly significant variables seems, nonetheless, to fail utterly on this criterion. It always predicts $y = 2$. By the

Count R^2 measure, our model achieves a fit of 0.4965, which looks like a substantial improvement over the *Pseudo* R^2 of 0.020785. Lest we become too enthusiastic about the result, however, note that the *Adjusted Count* R^2 is zero! The reason is that the model does not improve on the model free "always predict 2," which happens to be the most frequent outcome.

The situation in which the model always predicts the same value is not uncommon. It takes a high correlation (in some general sense) between the covariates and the outcome and a large amount of variation in the covariates within the sample to spread the predictions across the outcomes. Briefly, another example is provided by a standard data set used by the authors of *Stata* to demonstrate the ordered choice model in their documentation. The "automobile data," (http://www.stata-press.com/data/r8/fullauto.dta) is used in [R] `oprobit` to model the 1977 repair records of 66 foreign and domestic cars. The variable *rep77* takes values *poor, fair, average, good* and *excellent*. The explanatory variables in the model are *foreign* (origin of manufacture), *length* (a proxy for size) and *mpg*. (The computations below were done with both *Stata* and *NLOGIT*, which obtained identical results.) The predictions produced by this model are listed below in Table 5.8. The McFadden *Pseudo* R^2 is 0.1321. The *Count* R^2 is (1+0+21+7+1)/66 = 0.454. The adjusted value is (30 – 27)/(66-27) = 0.077.

Table 5.8 Predicted vs. Actual Outcomes for Automobile Data

Cross tabulation of predictions.						
Row is actual, column is predicted.						
Model=Probit. Prediction=most likely cell.						
Actual	0	1	2	3	4	Row sum
0	1	0	2	0	0	3
1	0	0	9	2	0	11
2	0	1	21	5	0	27
3	0	0	11	7	2	20
4	0	0	2	2	1	5
Col Sum	1	1	45	16	3	66

This survey does not conclude with a proposal for *the* appropriate or optimal fit measure. The search for a scalar counterpart to the R^2 in a linear regression does seem unproductive. Fit measures based on the log likelihood can be used for comparing models. For this purpose, the log likelihood itself or one of the information criteria seems sensible; the AIC dominates the received applications. For assessing the predictions of the model, it would seem that the scalar measures based on the log likelihood would be useless. The maximum likelihood estimator is not computed so as to maximise the number of correction predictions – in the linear normal regression model, the MLE of $\boldsymbol{\beta}$ is computed to maximize R^2, but that is coincidental; minimizing $\mathbf{e}'\mathbf{e}$ does maximize R^2. Indeed, there may be (as yet not proposed) other estimators that improve on the MLE for predicting the outcome variable, as the Maximum Score Estimator [see Manski (1975, 1985, 1986, 1988)] improves on the MLE of the logit or probit model for binary choice. In any event, it does seem appropriate, if one seeks a "measure of fit" one should first decide upon a procedure (rule) for producing the predictions, then assess, against a benchmark, how well that method does. The *Count* R^2 measures shown above seem better suited to that specific purpose than pseudo R^2 measures based on the log likelihood.

5.9 Estimation Issues

McKelvey and Zavoina (1975) provide expressions for the first and second derivatives of the log likelihood function for the ordered probit model, and suggest Newton's method as an algorithm for estimation. They do conjecture, however, about the possible problem of multiple roots of the log likelihood. Pratt (1981), was able to show that the ordered probit model was a member of a class of discrete choice models in which the log likelihood functions are globally concave. Thus, estimation of the model can be counted on to converge (when it does at all), to the single root of the log likelihood function. We note at this point a few other aspects of estimation of the ordered choice model.

5.9.1 Grouped Data

Grouped data arise when groups or sets of individuals have the same $\mathbf{x}_i$ and the observed outcome consists of a set of proportions over the choices. For example, in a taste test for a soft drink, $\mathbf{x}_i$ might consist of a specific configuration of (*sweetness*,*color*,*temperature*). A group of n_i individuals are presented with $\mathbf{x}_i$, and proportions p_{i0}, $p_{i1},\ldots,p_{iJ}$ of the n_i individuals choose outcome i. Thus, the frequency of individuals in group i reporting outcome j is $n_i \times p_{ij}$. In the boiassay experiments discussed in Chapter 4, x_i would the the dosage of insecticide administered to a group of n_i pests, and proportions p_{i0}, p_{i1} and p_{i2} are found to respond to the dosage by surviving, becoming moribund, or dying, respectively.

The adaptation of the maximum likelihood estimator to the grouped data treatment is a trivial modification. The log likelihood for a sample in which the stimulus, $\mathbf{x}_i$ is repeated n_i times is

$$\begin{aligned}
\log L &= \sum_{i=1}^{N}\sum_{t=1}^{n_i}\sum_{j=0}^{J} m_{it,j}\log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\right] \\
&= \sum_{i=1}^{N}\sum_{j=0}^{J}\log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\right]\sum_{t=1}^{n_i} m_{it,j} \\
&= \sum_{i=1}^{N}\sum_{j=0}^{J} n_{ij}\log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\right] \\
&= \sum_{i=1}^{N} n_i \sum_{j=0}^{J} p_{ij}\log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\right].
\end{aligned} \tag{5.12}$$

Mechanically, in the log likelihood for a cross section of individual data, the terms m_{ij} are replaced with the group proportions, p_{ij}, and the observations in the log likelihood and its derivatives are weighted by the group size.

5.9.2 Perfect Prediction

A problem of nonconvergence can be caused by a condition in the data that Long and Freese (2006, p. 192) label "Predicting Perfectly." If a variable in the data set predicts perfectly one of the implicit dependent variables, that is, $m_{ij} = 1$ if and only if $y_i = j$, then it will not be possible to fit the coefficients of the model – in this instance, the corresponding threshold parameter becomes inestimable. The suggested case is a dummy variable that takes only one value within a particular cell – it may also take that value in other cells. Within our example, suppose married people (*Married* = 1) always responded with *Health* = 4; i.e., married people always report the highest health satisfaction. Then, knowing someone is married allows a perfect prediction of *Health* = 4 for them. In such as case, it is necessary to drop such observations from the sample. *Stata* detects this condition automatically and reports a diagnostic "`Note: nn observations completely determined. Standard errors are questionable.`" As it is, the diagnostic is correct. But, it is incomplete. Because the offending variable enjoys

such a relationship with the outcome variable, it is almost certainly endogenous in the model, and not only are the standard errors questionable, the parameter estimates themselves are as well. In a vague way, this is a cousin to a problem of sample selection. The observations that have been discarded have not been done so randomly. They have been discarded by a criterion that is specifically related to the dependent variable. This particular feature of the model is as of this writing an obscure corner of the model development, but there would seem to be scope for further analysis of the issue.

It is tempting in this instance just to drop the offending variable. Whether this is advisable or not is unclear. If one is certain that but for the (perhaps unexpected) data problem the variable is an important feature of the data generating process, then the resulting model when the variable is dropped now has an omitted regressor. One problem has been traded for another. On the other hand, if the problem considered here involves more than just a handful of observations, one might question the overall structure of the model. Treating such a variable as if it were exogenous might be inappropriate.

5.9.3 Different Normalizations

We have noted at a few points that the normalization of the thresholds is a crucial feature of the model. However, it is not the case that different normalizations produce different results. Whether one assumes $\mu_0 = 0$ and includes an overall constant in the model, or allows μ_0 to be a free parameter and drops the constant, will have no implications for the log likelihood, the other parameters, or the predictions of the model. An example to illustrate the point is useful. Consider, once again, the car repair data discussed in the previous section. We have fit the model using *NLOGIT*, which uses the first normalization and *Stata* which uses the second. The two sets of results are given in Table 5.9. Note that the log likelihoods and estimates of the coefficients in $\boldsymbol{\beta}$ are identical. (The differences in the standard errors result from *Stata*'s use of the Hessian for the standard errors vs. *NLOGIT*'s use of the outer products estimator.) The first "cut point" in the *Stata* results is precisely the negative of *NLOGIT*'s overall constant. For the remaining threshold parameters, we can see that "cut point *j*" equals *NLOGIT's* ($\mu_j - \alpha$). As expected, then, the results are identical.

5.9.4 Censoring of the Dependent Variable

In some applications, there can be a second layer of censoring of the variable of interest in the ordered choice model. (The first level of censoring is the translation of y_i^* to y_i by measuring only the interval in which y_i^* appears.) Consider a model of educational attainment in which the variable of interest is "education" and in which the recorded value is only 0 for primary school, 1 for secondary school (high school), 2 for college, 3 for masters and 4 for Ph.D. If an observation is recorded as "at least high school," for example, then values 2, 3 and 4 are censored. This case is easily handled using the laws of probability. The appropriate log likelihood for the ordered choice model is

$$\log L = \sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \log(P_{ij} - P_{i,j-1}) , \tag{5.13}$$

where heretofore m_{ij} indicated the one cell that applies to observation i, and now indicates all of the cells that apply. For the example given, we would have $m_{i0} = 0$ and $m_{ij} = 1$ for $j = 1,2,3,4$. The change in the computations of the model parameters is trivial. It should be noted, one must know the upper bound, J, and for an observation, of course, it must be known that it is or is not censored. Censoring of the dependent variable in an ordered choice context has appeared in models of schooling attainment by Lillard and King (1987), Glewwe (1997) and Glewwe and

Jacoby (1994, 1995) and in duration models, where the observed outcome is the length of time between transitions, sometimes coded as "short," medium or long, or similarly. See, e.g., Tsay (2005), Han and Hausman (1988) and Buckle and Carlson (2000).

Table 5.9 Stata and NLOGIT Estimates of an Ordered Probit Model

```
. oprobit rep77 foreign length mpg
Iteration 0:   log likelihood = -89.895098
Iteration 1:   log likelihood = -78.141221
Iteration 2:   log likelihood = -78.020314
Iteration 3:   log likelihood = -78.020025
Ordered probit regression                         Number of obs   =         66
                                                  LR chi2(3)      =      23.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -78.020025                       Pseudo R2       =     0.1321
------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   1.704861   .4246786     4.01   0.000     .8725057    2.537215
      length |   .0468675    .012648     3.71   0.000      .022078    .0716571
         mpg |   .1304559   .0378627     3.45   0.001     .0562464    .2046654
-------------+----------------------------------------------------------------
       /cut1 |    10.1589   3.076749                      4.128586    16.18922
       /cut2 |   11.21003   3.107522                      5.119399    17.30066
       /cut3 |   12.54561   3.155228                      6.361476    18.72974
       /cut4 |   13.98059   3.218786                      7.671888     20.2893
------------------------------------------------------------------------------
Skip $ (The data on rep77 contain 8 missing observations)
Ordered Probit ; Lhs = rep77 ; Rhs=one,foreign,length,mpg $
+---------------------------------------------+
| Ordered Probability Model                   |
| Dependent variable               REP77      |
| Number of observations              66      |
| Log likelihood function       -78.02002     |
| Number of parameters                7       |
| Info. Criterion: AIC =          2.57636     |
| Restricted log likelihood     -89.89510     |
| McFadden Pseudo R-squared      .1320992     |
| Chi squared                   23.75015      |
| Degrees of freedom                  3       |
| Prob[ChiSqd > value] =         .2816655E-04 |
| Underlying probabilities based on Normal    |
+---------------------------------------------+
+--------+--------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |b/St.Er.|P[|Z|>z]| Mean of X|
+--------+--------------+----------------+--------+--------+----------+
---------+Index function for probability
 Constant|   -10.1589039      3.03379286    -3.349   .0008
 FOREIGN |    1.70486053       .41520516     4.106   .0000    .31818182
 LENGTH  |     .04686753       .01228262     3.816   .0001   189.121212
 MPG     |     .13045591       .03696460     3.529   .0004   21.3333333
---------+Threshold parameters for index
 Mu(1)   |    1.05112609       .18720281     5.615   .0000
 Mu(2)   |    2.38670648       .18420739    12.957   .0000
 Mu(3)   |    3.82169002       .28935433    13.208   .0000
```

5.9.5 Maximum Likelihood Estimation of the Ordered Choice Model

The log likelihood function for the basic ordered choice model is

$$\begin{aligned}\log L &= \sum_{i=1}^{n} n_i \sum_{j=0}^{J} w_{ij} \log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\right] \\ &= \sum_{i=1}^{n} n_i \sum_{j=0}^{J} w_{ij} \log(F_{i,j} - F_{i,j-1}) \\ &= \sum_{i=1}^{n} n_i \sum_{j=0}^{J} w_{ij} \log P_{i,j},\end{aligned} \tag{5.14}$$

where

n_i = the group size in the grouped data (typical bioassay) case or
$n_i = 1$ in the individual data case,

and

$w_{ij} = p_{ij}$ = the proportion of group i that responds with outcome j, or
$w_{ij} = m_{ij} = 1$ if individual i chooses outcome j in the individual data case.

$F(t)$ is the functional form in use, typically $\Lambda(t)$ for the ordered logit model or $\Phi(t)$ for the ordered probit model. For the moment, we will leave the functional form indeterminate. For obtaining the log likelihood and its derivatives, only the term $\log P_{i,j}$ is of consequence. The relevant derivatives are

$$\begin{aligned}&\frac{\partial \log P_{i,j}}{\partial \boldsymbol{\beta}} = \frac{f_{i,j} - f_{i,j-1}}{P_{i,j}}(-\mathbf{x}_i), \\ &\frac{\partial \log P_{i,j}}{\partial \mu_j} = \frac{f_{i,j}}{P_{i,j}}, \quad \frac{\partial \log P_{i,j}}{\partial \mu_{j-1}} = \frac{-f_{i,j-1}}{P_{i,j}},\end{aligned} \tag{5.15}$$

where $f_{i,j}$ is the density corresponding to $F_{i,j}$. For the moment, we are carrying μ_{-1}, μ_0 and μ_J as if they were unconstrained. The constraints are imposed later. Thus, the parameter vector contains $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, which has J+2 elements only J-1 of which are free to vary. The derivative vector $\partial \log P_{i,j}/\partial \boldsymbol{\mu}$ has J+2 elements, but only two are nonzero. The second derivatives are as follows:

$$\begin{aligned}&\frac{\partial^2 \log P_{i,j}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \left[\left(\frac{f'_{i,j} - f'_{i,j-1}}{P_{i,j}}\right) - \left(\frac{f_{i,j} - f_{i,j-1}}{P_{i,j}}\right)^2\right]\mathbf{x}_i\mathbf{x}'_i, \\ &\frac{\partial^2 \log P_{i,j}}{\partial \boldsymbol{\beta} \partial \mu_j} = \left[\frac{f'_{i,j}}{P_{i,j}} - \frac{(f_{i,j} - f_{i,j-1})f_{i,j}}{P_{i,j}^2}\right](-\mathbf{x}_i), \quad \frac{\partial^2 \log P_{i,j}}{\partial \boldsymbol{\beta} \partial \mu_{j-1}} = \left[\frac{-f'_{i,j-1}}{P_{i,j}} - \frac{(f_{i,j} - f_{i,j-1})(-f_{i,j-1})}{P_{i,j}^2}\right](-\mathbf{x}_i), \\ &\frac{\partial^2 \log P_{i,j}}{\partial \mu_j^2} = \left[\frac{f'_{i,j}}{P_{i,j}} - \left(\frac{f_{i,j}}{P_{i,j}}\right)^2\right], \quad \frac{\partial^2 \log P_{i,j}}{\partial \mu_{j-1}^2} = \left[\frac{-f'_{i,j-1}}{P_{i,j}} - \left(\frac{(-f_{i,j-1})}{P_{i,j}}\right)^2\right], \\ &\frac{\partial^2 \log P_{i,j}}{\partial \mu_j \partial \mu_{j-1}} = \left[\frac{(-f_{i,j})(-f_{i,j-1})}{P_{i,j}^2}\right].\end{aligned} \tag{5.16}$$

The Hessian has a nonzero 2×2 block within the full (J+2)×(J+2) submatrix for $\boldsymbol{\mu}$. The relevant constraints on the terms for the fixed elements of $\boldsymbol{\mu}$ are

$$\mu_{-1} = -\infty,\ \mu_0 = 0,\ \mu_J = \infty,$$
$$F_{i,-1} = 0,\ f_{i,-1} = 0,\ f_{i,-1}' = 0,$$
$$F_{i,J} = 1,\ f_{i,J} = 0,\ f_{i,J}' = 0.$$

Finally, for the two most commonly used functional forms,

$$\begin{aligned}
&\text{logit:} \quad F(t) = \Lambda(t),\\
&\qquad f(t) = \Lambda(t)[(1 - \Lambda(t)],\\
&\qquad f'(t) = \Lambda(t)[(1 - \Lambda(t)]\,[1 - 2\Lambda(t)],\\
&\text{probit:} \quad F(t) = \Phi(t),\\
&\qquad f(t) = \phi(t),\\
&\qquad f'(t) = -t\,\phi(t).
\end{aligned} \tag{5.17}$$

As Pratt (1981) showed, the second derivatives matrix is negative definite, so common gradient methods such as Newton or BFGS should be effective for maximizing the log likelihood function. Occasionally (rarely in our experience, however), the threshold parameters can become unordered during optimization. This points to the utility of a line search and a careful iteration. It is possible to force the threshold parameters to be ordered by reparameterizing them. For the model proposed in Section 8.3, we used the formulation

$$\mu_j = \mu_{j-1} + \exp(\alpha_j).$$

starting with $\mu_0 = 0$.

5.9.6 Bayesian (MCMC) Estimation of Ordered Choice Models

Bayesian estimation of ordered choice models builds on the method pioneered by Albert and Chib (1993). The Gibbs sampler is constructed using a crucial device labeled "data augmentation." [See Tanner and Wong (1987).] The binary choice case departs from

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim \text{ with mean 0 and known variance, 1 (probit) or } \pi^2/3 \text{ (logit)},$$

$$y_i = 1 \text{ if } y_i^* > 0.$$

Let the prior for $\boldsymbol{\beta}$ be denoted $p(\boldsymbol{\beta})$. Then, the posterior density for the probit or logit (symmetric distribution) models is

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \frac{p(\boldsymbol{\beta})\prod_{i=1}^{n} F[(2y_i - 1)\boldsymbol{\beta}'\mathbf{x}_i]}{\int_{\boldsymbol{\beta}} p(\boldsymbol{\beta})\prod_{i=1}^{n} F[(2y_i - 1)\boldsymbol{\beta}'\mathbf{x}_i]d\boldsymbol{\beta}}, \tag{5.18}$$

where we use $\mathbf{y}$ and $\mathbf{X}$ (and later, $\mathbf{y}^*$) to denote the full set of n observations. [See (2.25).] Estimation of the posterior mean is done by setting up a Gibbs sampler in which the unknown values y_i^* are treated as nuisance parameters to be estimated. For convenience at this point, we will assume the probit model is of interest. Conditioned on $\boldsymbol{\beta}$ and $\mathbf{x}_i$, y_i^* has a normal distribution with mean $\boldsymbol{\beta}'\mathbf{x}_i$ and variance 1. However, when conditioned on y_i (observed), as well, the sign of y_i^* is known;

$p(y_i^* \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ = normal with mean $\boldsymbol{\beta}'\mathbf{x}_i$ and variance 1, truncated at zero; truncated from below if $y_i = 1$ and from above if $y_i = 0$.

Using basic results for Bayesian analysis of the linear model with known disturbance [see Greene (2008a, p. 605)] and a diffuse prior, the posterior for $\boldsymbol{\beta}$ conditioned on $\mathbf{y}^*$, $\mathbf{y}$ and $\mathbf{X}$ would be

$$p(\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{y}, \mathbf{X}) = N_K[\mathbf{b}, (\mathbf{X}'\mathbf{X})^{-1}] \text{ where } \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*.$$

If, instead, the prior for $\boldsymbol{\beta}$ is normal with mean $\boldsymbol{\beta}^0$ and covariance matrix, $\boldsymbol{\Sigma}$, then the posterior density is normal with mean

$$E[\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{y}, \mathbf{X}] = [\boldsymbol{\Sigma}^{-1} + (\mathbf{X}'\mathbf{X})]^{-1} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}^0 + \mathbf{X}'\mathbf{y}^*)$$

and

$$\text{Var}[\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{y}, \mathbf{X}] = [\boldsymbol{\Sigma}^{-1} + (\mathbf{X}'\mathbf{X})]^{-1}.$$

This sets up a strikingly simple Gibbs sampler for drawing from the joint posterior, $p(\boldsymbol{\beta}, \mathbf{y}^* \mid \mathbf{y}, \mathbf{X})$. It is customary to use a diffuse prior for $\boldsymbol{\beta}$. Then, compute initially, $(\mathbf{X}'\mathbf{X})^{-1}$ and the lower triangular Cholesky matrix, $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}' = (\mathbf{X}'\mathbf{X})^{-1}$. (The matrix $\mathbf{L}$ needs only to be computed only once at the outset for the informative prior as well.) To initialize the iterations, any reasonable value of $\boldsymbol{\beta}$ may be used. Albert and Chib suggest the classical MLE. The iterations are then given by

1. Compute the N draws from $p(\mathbf{y}^* \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$.
   Draws from the appropriate truncated normal can be obtained using

$$y_i^*(r) = \boldsymbol{\beta}'\mathbf{x}_i + \Phi^{-1}[\Phi(-\boldsymbol{\beta}'\mathbf{x}_i) + U(1-\Phi(-\boldsymbol{\beta}'\mathbf{x}_i))] \text{ if } y_i = 1 \text{ and}$$

$$y_i^*\text{“}(r) = \boldsymbol{\beta}'\mathbf{x}_i + \Phi^{-1}[U\,\Phi(-\boldsymbol{\beta}'\mathbf{x}_i)] \text{ if } y_i = 0, \qquad (5.19)$$

   where U is a single draw from a standard uniform population.
2. Draw an observation on $\boldsymbol{\beta}$ from the posterior $p(\boldsymbol{\beta} \mid \mathbf{y}^*, \mathbf{y}, \mathbf{X})$ by first computing the mean

$$\mathbf{b}\text{“}(r) = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y}^*\text{®}.$$

   Use a draw, $\mathbf{v}$, from the K-variate standard normal, then compute $\boldsymbol{\beta}\text{“}(r) = \mathbf{b}\text{“}(r) + \mathbf{L}\mathbf{v}$.

(We have used “(r)” to denote the rth cycle of the iteration.) The iteration cycles between steps 1 and 2 until a satisfactory number of draws is obtained (and a burn-in number are discarded), then the retained observations on $\boldsymbol{\beta}$ are analyzed. With an informative prior, the draws at step 2 involving the prior mean and variance are slightly more time consuming. The matrix $\mathbf{L}$ is only computed at the outset, but the computation of the mean adds a matrix multiplication and addition.

The extension to J+1 ordered outcomes is now straightforward. We maintain the probit model, as is common. The model is, now,

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \; \varepsilon_i \sim N[0, 1],$$

$$y_i = j \text{ if } \mu_{j-1} < y_i^* < \mu_j.$$

Diffuse priors are assumed for $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, with the usual constraints on μ_{-1}, μ_0 and μ_J. Based on the same results as before, we still have

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}^*, \boldsymbol{\mu}, \mathbf{y}, \mathbf{X}) &= N_K[\mathbf{b}, (\mathbf{X}'\mathbf{X})^{-1}]. \\ p(y_i^* \mid \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) &= N(\boldsymbol{\beta}'\mathbf{x}_i, 1) \text{ truncated in both tails by } \mu_{j-1} \text{ and } \mu_j. \end{aligned} \tag{5.20}$$

We will note below how to do the simulation for y_i^*. Finally, the authors provide the posterior for μ_j (j = 1,...,J-1), conditioned on the other threshold parameters,;

$$p(\mu_j \mid \boldsymbol{\beta}, \mathbf{y}^*, \boldsymbol{\mu}_{(j)}, \mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^{n} \left\{ \begin{matrix} 1[y_i = j] \times 1[\mu_{j-1} < y_i^* < \mu_j] + \\ 1[y_i = j+1] \times 1[\mu_j < y_i^* < \mu_{j+1}] \end{matrix} \right\}, \tag{5.21}$$

where the density is the posterior for μ_j given the other threshold parameters, denoted $\mu_{(j)}$, and the other parameters. The steps in the Gibbs sampler consist of initializing $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ as before, now with the MLE of the ordered probit model, then, in order,

1. Sample μ_j from a uniform distribution with limits

   $Lower = \max_i\{\max(y_i^* | y_i = j), \mu_{j-1}\}$ (i.e., the maximum over the n observations),

   $Upper = \min_i\{\min(y_i^* | y_i = j+1), \mu_{j+1}\}$.

   Sampling from this uniform distribution is easily done by scaling a draw from $U(0,1)$ by $1/(Upper - Lower)$.

2. Sample y_i^* from the truncated normal distribution where the underlying variable has mean $\boldsymbol{\beta}'\mathbf{x}_i$ and standard deviation 1 and the truncation limits are μ_{j-1} and μ_j for the corresponding observation on $y_i = j$. The necessary result for this step is given in Greene (2008a, p. 575). To sample a draw from this distribution, define $P_L = \Phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)$ and $P_U = \Phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)$. Note that $P_L = 0$ if $y_i = 0$, and $P_U = 1$ if $y_i = J$. Then, let U denote a draw from the $U(0,1)$ population – a single uniform draw. Then, the draw for y_i^* is

   $$y_i^* | y_i, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{x}_i = \boldsymbol{\beta}'\mathbf{x}_i + \Phi^{-1}[P_L + U \times (P_U - P_L)].$$

3. Sample $\boldsymbol{\beta}$ from the multivariate normal population as shown earlier for the binary probit case. The only change is the data used to compute $\mathbf{b}$, now using the results of the doubly truncated sample in step 2 immediately above.

We then cycle through steps 1 – 3 for a large number of iterations (say tens of thousands). After discarding the first several thousand draws, the remaining draws on $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ constitute a sample from the joint posterior. The posterior mean is estimated by the average of the draws.

A convenient aspect of the MCMC approach to estimation is that often the estimator for a more complex model is easily obtained by adding layers to a simpler one. Consider the bivariate ordered probit model analyzed by Biswas and Das (2002). The model is a direct extension of the univariate model:

$$
\begin{aligned}
y_{i1}^* &= \boldsymbol{\beta}_1'\mathbf{x}_{i1} + \varepsilon_{i1},\ \varepsilon_{i1} \sim N[0, 1],\\
y_{i1} &= j \text{ if } \mu_{j-1} < y_{i1}^* < \mu_j\\
y_{i2}^* &= \boldsymbol{\beta}_2'\mathbf{x}_{i2} + \varepsilon_{i2},\ \varepsilon_{i2} \sim N[0, 1],\\
y_{i2} &= k \text{ if } \gamma_{k-1} < y_{i2}^* < \gamma_k.\\
&\text{Corr}(\varepsilon_{i1},\varepsilon_{i2}) = \rho.
\end{aligned}
$$

Each ordered probit is handled as before. The draws from the posterior of $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ are obtained by a two equation GLS regression; conditioned on the other parameters, the two latent regressions are a seemingly unrelated regressions system. The draws for (μ_j,γ_k) are drawn jointly from a rectangle, with each dimension handled as in the univariate case. The draws on y_{i1}^* and y_{i2}^* are drawn from a truncated bivariate normal population. (Biswas and Das suggest to do this draw by a rejection method. It can be done in a "one draw" manner using a bivariate truncated normal analog to the method shown above. [See, e.g., Geweke (1991).]) The remaining detail is sampling from the posterior of ρ. Biswas and Das handle this by defining $\boldsymbol{\Sigma}$ to be an *unrestricted* 2×2 covariance matrix of the two disturbances. The prior for $\boldsymbol{\Sigma}$ is assumed to be proportional to $|\boldsymbol{\Sigma}|^{-3/2}$. This produces a conditional posterior for $\boldsymbol{\Sigma}$ that is an inverse Wishart population. [See Train (2003) for sampling from this population.] Note that they have introduced two new free parameters, σ_{11} and σ_{22} and are now estimating $\sigma_{12} = \rho\sigma_1\sigma_2$.

There is a peculiar loose end in the Biswas and Das (2002) study. In the ordered choice model, the scale parameters of the disturbances, $\sigma_m^2 = \text{Var}[\varepsilon_{im}]$ are not identified and are normalized to 1.0. (In an alternative normalization of the model, one of the slopes is normalized at 1.0, which "identifies" the scale parameter – though not actually if that scale parameter is meant to be interpreted as the variation of ε. It merely moves the normalization off one of the parameters. See Chapter 12 for applications.) Biswas and Das treated these variances as free parameters, and did not normalize one of the other parameters. As such, the model they purport to estimate is not identified. The evidence is in the reported values of the posterior means of $\sigma_1^2 = 22.62$ and $\sigma_2^2 = 13.33$. These values are far outside the reasonable range for a choice model of this sort; they are supposed to be normalized at 1.0. (One might surmise that they are "identified" purely by the prior; there is no sample information about them.) This application points up a note of caution needed in MCMC estimation. The log likelihood function developed in Section 10.2 cannot be maximized it if is formulated in terms of an unrestricted $\boldsymbol{\Sigma}$ as used above. Ultimately, the derivatives will be collinear and the Hessian will be singular – that is the impact of a model that contains unidentified parameters. There is no counterpart control when using the Gibbs sampler. The signal that something has gone awry will arrive when the chain fails to converge, or when it arrives at a very different vector of posterior means from one run to another. It is necessary to check these failures – one run of the Gibbs sampler, regardless of how long it is, will not reveal this condition. (Redemption of the model would be obtained by formulating it in terms of a prior over ρ to begin with, and imposing the necessary normalizations on σ_1 and σ_2.)

As noted earlier, the Bayesian segment of this literature is relatively compact and quite recent. Methodological contributions are offered by Albert and Chib (1993), Koop and Tobias (2006) and Imai et al. (2003) who have developed an "R" routine for some of the computations. Applications include Girard and Parent (2001), Biswas and Das (2002), Czado, Heyn and Müller (2005), Tomoyuki et al. (2006), Ando (2006), Zhang et al. (2007), Kadam and Lenk (2008) and Munkin and Trivedi (2008) and a handful of others. Doubtless there are more to come. Nonetheless, as of this writing, Bayesian analysis of ordered choice data is a small niche in the literature. There are, of course, a cornucopia of applications to binary data.

5.9.7 Software For Estimation of Ordered Choice Models

There are numerous commercial packages that can be used to estimate basic ordered choice models. (We mention the packages only by name here. Each of them is described in detail on their own respective website, listed in Table 5.10, so we will forego any detailed descriptions.) The primary ones in current use are *SAS*, *Stata*, *LIMDEP*, *NLOGIT* and *SPSS*. In addition, *Latent Gold* and a few other programs less oriented to cross section and panel data, such as *RATS*, *Eviews* and *TSP*, also contain built-in estimators for the essential model. For Bayesians, there are routines in *R* provided in *ZELIG* by Imai et al. (2008). *WinBugs* also contains a routine for discrete choice models. The log likelihood is not particularly complicated, and *Gauss* and *Matlab* programs are also widely circulated.

For more advanced, exotic or obscure variants of the model, the choices are much more limited. These can, of course, be programmed by the user in the low level languages such as *Matlab*, or in many cases, even in the higher level matrix languages of the integrated packages such as *Stata*. For prepackaged routines, *Stata* and *NLOGIT/LIMDEP* contain optional features, such as heteroscedasticity and individual specific thresholds. Models with random coefficients can be fit with *PROC MIXED* in *SAS*, `GLAMM` in *Stata*, and with several of the routines in *NLOGIT*. To our knowledge, only *Latent Gold* and *NLOGIT/LIMDEP* have built in latent class treatments for ordered choice models. For panel data applications, the random effects model (Butler and Moffitt) is quite common as well and appears in all the familiar packages. Random effects models are "random constants" models. So any random parameters module can also handle random effects in a panel. That we are aware of, the fixed effects model with essentially unlimited numbers of effects (beyond the capacity to just add the dummy variables to the model) is available only in *NLOGIT and LIMDEP*.

The following is a list of the websites of the packages mentioned above. This is far from a complete list of software used in econometrics and statistics. For a lengthy guide that comes close to one, the econometric software resource

Econometrics `http://www.oswego.edu/~economic/econsoftware.htm`

is a useful reference point. The widely used packages are listed in Table 5.10:

Table 5.10 Software Used for Ordered Choice Modeling

Eviews	`http://www.eviews.com`
Gauss	`http://www.aptech.com`
Latent Gold	`http://www.statisticalinnovations.com/`
LIMDEP	`http://www.limdep.com`
Matlab	`http://www.mathworks.com`
NLOGIT	`http://www.nlogit.com`
RATS	`http://www.estima.com`
SAS	`http://www.sas.com`
SPSS	`http://www.spss.com`
Stata	`http://www.stata.com`
TSP	`http://www.tspintl.com`
WinBugs	`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`
ZELIG	`http://gking.harvard.edu/zelig/`

6

Specification Issues and Generalized Models

Anderson (1984, p. 2) discusses the inadequacy of the ordered choice model we have examined thus far. "We argue here that the class of regression models currently available for ordered categorical response variables is not wide enough to cover the range of problems that arise in practice. Factors affecting the kind of regression model required are (i) the type of ordered categorical variable, (ii) the observer error process and (iii) the "dimensionality" of the regression relationship. These factors relate to the processes giving rise to the observations and have been rather neglected in the literature." Generalizations of the model, e.g., Williams (2006), have been predicated on Anderson's observations, as well as some observed peculiarities in data being analyzed.

It is useful to distinguish between two directions of the contemporary development of the ordered choice model. Although it hints at some subtle aspects of the model (underlying data generating process), Anderson's arguments, it will emerge, direct attention to the functional form of the model and its inadequacy in certain situations. Beginning with Terza (1985), a number of authors have focused, instead, on the fact that the model does not account adequately for individual heterogeneity that is likely to be present in micro- level data. A series of themes are addressed in this chapter as we build up to the most general ordered choice model.

6.1 Functional Form Issues and the Generalized Ordered Choice Model (1)

Once again, referring to Anderson (1984, p. 2), "The dimensionality of the regression relationship between y and $\mathbf{x}$ is determined by the number of linear functions required to describe the relationship. If only one linear function is required, the relationship is one-dimensional; otherwise it is multi-dimensional. For example, in predicting k categories of pain relief from predictors $\mathbf{x}$, suppose that different functions $\boldsymbol{\beta}_1'\mathbf{x}$ and $\boldsymbol{\beta}_2'\mathbf{x}$ are required to distinguish between the pairs of categories (*worse, same*) and (*same, better*), respectively. Then, the relationship is neither one-dimensional nor ordered with respect to $\mathbf{x}$." The fundamental flaw in the argument is in its opening premise. There is no regression relationship between y and $\mathbf{x}$. The observed variable is merely a set of labels. What follows is curve fitting – suggesting that two equations might better fit two binary choices than a single one. (It remains to determine by what criterion different functions are *required*.) On the other hand, the author's earlier (also p. 2) analysis of the data generating process puts a better face on the argument.

> For example, Anderson and Philips (1981) refer to the "extent of pain relief after treatment:" *worse*, *same*, *slight improvement*, *moderate improvement*, *marked improvement* or *complete relief. In principle, there is a single, unobservable, continuous variable related to this ordered scale,* [emphasis added] but in practice, the doctor making the assessment will use several pieces of information in making his judgment on the observed category. For example, he might use severity of pain, kind of pain, consistency in the time and degree of disability. We will refer to variables of the second type as "assessed" ordered categorical variables and argue that, in general, a different approach to modeling regression relationships is appropriate for the two types. Assessed ordered variables occur frequently in the biomedical, social and other social sciences.

Thus, he argues that, at least in some situations, the dependent variable is not really ordered, or might not be. In such a case, he argues, essentially, that it makes sense to partition the outcomes, and treat them as a set of binary choices, or at least not as a single ordered choice. For the

specific application considered, the issue depends crucially on whose assessment is being recorded, the doctor's (not necessarily cleanly ordered as measured against some objective yardstick) or the patient's (one would assume, necessarily ordered). The upshot is that, at least as argued here, increasng the "dimensionality" of the fitting problem follows from the nature of the data generating process, not (evidently) from a need to accommodate curvature in the data.

6.1.1 Parallel Regressions

Anderson departs from the familiar ordered choice model that we have examined so far;

$$\text{Prob}(y \leq y_s \mid \mathbf{x}) = F(\theta_s - \boldsymbol{\beta}'\mathbf{x}),\ s = 1,...,k.$$

Continuing the line of argument suggested earlier, he then suggests his "new" model,

$$\text{Prob}(y = y_s \mid \mathbf{x}) = \frac{\exp(\beta_{0s}^* + \boldsymbol{\beta}_s'\mathbf{x})}{\Sigma_{t=0}^{k} \exp(\beta_{0t}^* + \boldsymbol{\beta}_t'\mathbf{x})},\ s = 1,...,k,\ \beta_{0k}^* = 0,\ \boldsymbol{\beta}_k = \mathbf{0}. \tag{6.1}$$

This is the multinomial logit model proposed by Nerlove and Press (1972) for *k unordered* choices. Later, it is observed "Model (5) [the model immediately above] often gives a good fit to real data, even when the $\boldsymbol{\beta}_s$ are *restricted to be parallel*. This is particularly true when the categories are ordered." [Emphasis added.] Thus appears (apparently) the first occurrence of the "parallel regressions" notion in this literature. Note the implication is that the model is not intended for ordered data; but it seems to work well when applied to ordered outcomes. By "parallel," the author states the restriction $\boldsymbol{\beta}_s = -\phi_s\boldsymbol{\beta}$ where $\phi_k \equiv 0$. [Note that the last ϕ_s is a parameter that is not identified under either the null or the alternative hypothesis because the corresponding $\boldsymbol{\beta}_s = 0$. See Andrews and Ploberger (1994).] A further identifying normalization (no longer merely for convenience) is $\phi_1 \equiv 1$. The resulting model,

$$\frac{\text{Pr}(y = y_s \mid \mathbf{x})}{\text{Pr}(y = y_k \mid \mathbf{x})} = \exp(\beta_{0s} - \phi_s\boldsymbol{\beta}), s = 1,...,k \text{ [8]}, \tag{6.2}$$

is labeled the "*Stereotype Ordered Regression Model*." As stated, the name is a misnomer, as the model does not enforce the ordering of the outcome; it is simply a parametric restriction on a model for *unordered* outcomes. [See Theil (1970).] Indeed, no linear restriction on the parameters of this model can enforce the ordering of the dependent variable, that is, the sequence

$$\text{Pr}(y \leq y_s|\mathbf{x}) < \text{Pr}(y \leq y_{s+1}|\mathbf{x}).$$

As he notes, the model "often gives a good fit to real data." However, the ordering aspect of it would depend on the data. It is not a feature of the model. We should note, the underlying structure has been lost in this process. It is not possible to discern what underlying data generating process would give rise to such a functional form for a strictly ordered outcome that arises from an underlying continuous measure.

Anderson follows with a prescription for enforcing the ordering of the outcomes. "The next step is to order the $\boldsymbol{\beta}_s$ to obtain a regression relationship. This is achieved by ordering the ϕ_s,

$$1 = \phi_1 > \phi_2 > ... > \phi_k = 0. \text{ [10]}.$$

"The ordered regression model [8] subject to constraints [10] will be termed the stereotype model." This form is prescribed for ordered data. Unfortunately, the model is still not out of difficulty. The implied probabilities still do not enforce the ordering rule unless the constant terms are monotonically increasing; $\beta_{01} < \beta_{02} < ... < \beta_{0k}$. Thus, Anderson's remedy for the "parallel regressions" restriction, if we enforce the ordering of the probabilities, is a progressive scaling of the parameter vector by the constants ϕ_s, but it is not an internally consistent model for ordered choices without the constraint on the constant terms.

Long (1997) departs from our (now) familiar formulation of the ordered choice model.

$$\text{Prob}(y \leq j \mid \mathbf{x}_i) = F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i).$$

Differentiating these functions, we have

$$\partial \text{Prob}[y_i \leq j \mid \mathbf{x}_i]/\partial \mathbf{x}_i = -f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)\,\boldsymbol{\beta}. \tag{6.3}$$

This defines a set of binary choice models with different constants but common slope vector, $\boldsymbol{\beta}$. If we then fix the probability at, say $P = P^*$ for any outcome, it must follow (by monotonicity of the cdf) that $f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)$ is fixed at f^*. It follows that *for a particular choice of probability*, we have

$$\partial \text{Prob}[y_i \leq j \mid \mathbf{x}_i]/\partial \mathbf{x}_i = f^*\,\boldsymbol{\beta} = \partial \text{Prob}[y_i \leq m \mid \mathbf{x}_i]/\partial \mathbf{x},\ m = 0,...,J. \tag{6.4}$$

where f^* is the same for all j, that is, a multiple of the same $\boldsymbol{\beta}$. This is the feature of the model that has been labeled the "parallel regression assumption." [See, e.g., Long (1997, p. 141).] This is an intrinsic feature of the ordered choice model. There is no obvious implication of the restriction for the underlying behavioral assumption – we will examine this issue in the next section. Note that the restriction cannot hold for a particular individual, since it requires the thresholds to adjust to equality. (I.e., we cannot fix all the probabilities to equal the chosen value at the same time. Rather, the "restriction" states that if P_1 equals P^*, then the derivative is the same as if P_2 equals the same P^*.)

6.1.2 Testing the Parallel Regressions Assumption – The Brant (1990) Test

Brant (1990), approaches the parallel regressions issue, but couches it in different terms. Defining

$$\gamma_j = \text{Prob}(y \leq j \mid \mathbf{x}) = F(\mu_j - \boldsymbol{\beta}'\mathbf{x}),$$

the logit form of the model implies (as well) that

$$\log\left(\frac{\gamma_j}{1-\gamma_j}\right) = \mu_j - \boldsymbol{\beta}'\mathbf{x}, \tag{6.5}$$

a "restriction" labeled the "proportional odds" restriction, or the "proportional odds model." [McCullagh (1980)] Brant notes, this *is* a testable restriction, as we explore shortly. One is left to wonder, what feature of the model, or of the behavior underlying it, has been revealed when the null "hypothesis" of parallel regressions is rejected statistically, as it frequently is. Other than the purely mechanical observation that in a "model" with different coefficient vectors for each choice, the parallel regressions restriction is that those coefficients are the same, it is unclear in

modeling terms, what the assumption means. Brant raised the same question. Before we reconsider that question, we will examine the proposed test procedure.

Several approaches to examining the parallel regressions feature have been developed. All center on the set of implied binary choice "models" for the probit and logit cases,

$$\text{Prob}(y \geq j \mid \mathbf{x}) = F(\boldsymbol{\beta}'\mathbf{x} - \mu_j), j = 1,...,J\text{-}1. \tag{6.6}$$

Thus, one can, in principle, fit J-1 such models separately. Each should produce its own constant term and a consistent estimator of the common $\boldsymbol{\beta}$. An "informal" examination of the differences [see Clogg and Shihadeh (1994, pp. 159-160)] should be revealing. A Lagrange multiplier test of the hypothesis is presented by SAS Institute (2008). A much more straightforward (and intuitive) test is Brant's (1990) Wald test which directly examines the restrictions

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = ... = \boldsymbol{\beta}_{J\text{-}1}.$$

The Brant (1990) test of this hypothesis for the ordered logit model follows from the implication of the model,

$$\text{Prob}[y_i \geq j \mid \mathbf{x}_i] = \Lambda(\beta_{0j} + \boldsymbol{\beta}_j'\mathbf{x}_i), \tag{6.7}$$

where $\beta_{0j} = \beta_0 - \mu_j$ and $\Lambda(t)$ is the logistic cdf, $1/(1+\exp(-t))$. The slope vector $\boldsymbol{\beta}_j$ should be the same in every equation. Thus, the specification implies J-1 binary choice "models" that can be estimated one at a time, each with its own constant term and (by assumption) the same slope vector.

Expressions for the mechanics of the test appear in Long (1997, pp. 144-145.) The null hypothesis is equivalent to

$$H_0: \boldsymbol{\beta}_q - \boldsymbol{\beta}_1 = \mathbf{0}, q = 2,...,J\text{-}1,$$

which can be summarized as

$$H_0: \mathbf{R}\boldsymbol{\beta}^* = \mathbf{0}$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{I} \end{bmatrix}, \quad \boldsymbol{\beta}^* = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \\ \vdots \\ \boldsymbol{\beta}_{J-1} \end{bmatrix}. \tag{6.8}$$

The Wald statistic will be

$$\chi^2[(J\text{-}1)K] = \left(\mathbf{R}\hat{\boldsymbol{\beta}}^*\right)'\left[\mathbf{R} \times \mathbf{Asy.Var}\left[\hat{\boldsymbol{\beta}}^*\right] \times \mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}^*\right), \tag{6.9}$$

where $\hat{\boldsymbol{\beta}}^*$ is obtained by stacking the individual binary logit estimates of $\boldsymbol{\beta}$ (without the constant terms). The remaining complication in (6.9) is the asymptotic covariance matrix, which is computed as follows (using Brant's results):

$$Est.Asy.Cov\left[\hat{\boldsymbol{\beta}}_j,\hat{\boldsymbol{\beta}}_m\right]=\left[\sum_{i=1}^{n}\hat{\Lambda}_{ij}\left(1-\hat{\Lambda}_{ij}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\times \left[\sum_{i=1}^{n}\hat{\Lambda}_{im}\left(1-\hat{\Lambda}_{ij}\right)\mathbf{x}_i\mathbf{x}_i'\right]\left[\sum_{i=1}^{n}\hat{\Lambda}_{im}\left(1-\hat{\Lambda}_{im}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \tag{6.10}$$

and $\hat{\Lambda}_{ij}=\Lambda\left(\hat{\beta}_{0j}+\hat{\boldsymbol{\beta}}_j'\mathbf{x}_i\right)$. The test can be carried out for specific coefficients by removing all but the desired rows of **R** in the computation of the statistic.

There are some loose ends in the computation. If the probabilities in the covariance matrix are based on the individual binary logit models, then the ordering of the probabilities is not preserved, and Λ_{ij} - $\Lambda_{i,j-1} < 0$ is a possibility even though the theory rules it out. Brant suggests using the parameters of the restricted (basic ordered choice) model instead. Even with this practical fix, it remains true that the parameter estimates used in the test, each of which does have its own constant term, do not preserve the ordering of the probabilities in the model.

Table 6.1 displays the results of the Brant test for our ordered logit model of health satisfaction. The proportional odds restriction is clearly rejected. Loosely, it appears that the income coefficient displays the greatest variation across the cells. Both education and income appear to fail the test when it is applied individually.

Table 6.1 Brant Test for Parameter Homogeneity

```
+-----------------------------------------------+
| Brant specification test for equal coefficient |
| vectors in the ordered logit model. The model  |
| implies that logit[Prob(y>j|x)]=beta(j)*x - mj |
| for all j = 0,..., 3. The chi squared test is  |
| H0:beta(0) = beta(1) = ... beta( 3)            |
| Chi squared test statistic =     71.76435      | (78.76988 based on the
| Degrees of freedom         =     15            | normal distribution)
| P value                    =       .00000      |
+-----------------------------------------------+
+-------------------------------------------------------------------------+
|Specification Tests for Individual Coefficients in Ordered Logit Model  |
|Degrees of freedom for each of these tests is  3                         |
+--------+-----------------+-------+-------+-------+-------+-------------+
|        |   Brant Test    | Coefficients in implied model Prob(y > j).  |
|Variable| Chi-sq  P value |   0   |   1   |   2   |   3   |             |
+--------+-----------------+-------+-------+-------+-------+-------------+
|AGE     |   6.28   .09864 | -.0398| -.0292| -.0328| -.0248|             |
|EDUC    |  19.89   .00018 |  .1212|  .0786|  .0630| -.0044|             |
|INCOME  |  13.32   .00398 | 1.9576|  .4959|  .1790| -.0206|             |
|MARRIED |   1.87   .59962 |  .0674| -.0228| -.1486| -.0896|             |
|KIDS    |   7.24   .06476 |  .3218|  .2158|  .0189| -.1231|             |
+--------+-----------------+-------+-------+-------+-------+-------------+
```

This naturally leads to some question of the model specification. For reasons we examine in more detail below, the non-proportional odds formulation is not a valid specification for the ordered logit model. Among the obvious reasons, the probabilities in the non-proportional odds model do not sum to one. If all the parameters can vary freely, as they do above, then each of the *J* binary choice models has been treated separately, and with no connection, there is no restriction on the sum of the probabilities. Moreover, there is no parametric restriction other than the one we seek to avoid that will preserve the ordering of the probabilities for all values of the data – that it does so for some data sets, or is a good "approximation" still leaves open the question of what specification failure makes sense to explain the finding, such as ours above.

Brant speculates at length about what model failures might lead to rejection of the hypothesis. The possibilities he lists include:

(1) Misspecification of the latent regression, $\boldsymbol{\beta}'\mathbf{x}$,
(2) Heteroscedasticity of ε - "nonhomogeneous dispersion of the latent variable with varying **x**."
(3) Misspecification of the distributional form for the latent variable, i.e., "nonlogistic link function."

He also considers a type of measurement error, such as the problem of "differential misclassification in the *y* observations." Brant expresses little optimism that the test will likely uncover failures (1) or (2), reasoning that if the index or the variance are misspecified in the structural model, the misspecification will distort the estimators in the binary choice models similarly. For the distributional assumption, however, he shows that if some other distribution applies, such as the extreme value distribution, then the appropriate model should echo something similar to Anderson's (1984) stereotype model, that is, with *j*-specific parameter vectors, $(\theta_j,\phi_j\boldsymbol{\beta})$. In this case, rejection of the common $\boldsymbol{\beta}$ form in favor of the more general form would be expected. Note, though that even under this assumption, this does not suggest that one should expect to find completely separate $\boldsymbol{\beta}_j$s. The differential multiple follows from the fact that even under the alternative distribution, the function is still parameterized in terms of a single index function. The scale factor is being induced by the different (from the logit) shape of the cdf with that same index function as its argument.

A more direct approach to testing against the distributional assumption is proposed by Johnson (1996) and Glewwe (1997). For this purpose, the null model is the ordered probit model based on the normal distribution. The Lagrange multiplier test is constructed by nesting the normal distribution within the broader Pearson family of distributions then testing against the null hypothesis of certain values of the parameters in the general form. [See Johnson, Kotz and Balakrishnan (1994).] It is noteworthy, at the end of the analysis, Glewwe (1997, p. 12) comes to the same juncture we have here. "A final question is what an applied econometrician should do when an ordered probit model does not pass the specification test." Like all specification tests, the "alternative" is not well defined. Glewwe surmises that the test might be picking up an altogether different failure, such as an incorrect functional form. He does suggest some alternative strategies, and ultimately suggests that if the failure of the LM test persists, perhaps an ordered logit might be preferable.

The Brant test is easily transported to the ordered probit model. Using the usual approximation, each maximum likelihood binary choice estimator converges to

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \mathbf{H}_j^{-1}\mathbf{g}_j + o(1/n),$$

where $\mathbf{H}_j^{-1}$ is the inverse of the information matrix and $\mathbf{g}_j$ is the gradient of the log likelihood. Relying on the information matrix equality and the results of Berndt, Hall, Hall and Hausman (1974), we can estimate the matrix using the outer product of gradients and estimate the covariances of the derivatives with the sum of cross products. For the binary probit models,

$$\mathbf{g}_{ij} = \frac{(2q_{ij}-1)\phi(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)}{\Phi[(2q_{ij}-1)\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i]}(\mathbf{x}_i),$$

where $q_{ij} = 1(y_i > j)$. The estimators of the submatrices needed for the test are

$$Est.Asy.Cov\left[\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_m\right] = \left[\sum_{i=1}^{n} \mathbf{g}_{ij}\mathbf{g}_{ij}'\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{g}_{ij}\mathbf{g}_{im}'\right]\left[\sum_{i=1}^{n} \mathbf{g}_{im}\mathbf{g}_{im}'\right]^{-1}. \quad (6.11)$$

Evidently this is not the explanation for the finding in Table 6.1. When we repeated the computations in Table 6.1 based on the ordered probit model, the chi squared statistic rose to 78.77.)

An intriguing point of the argument here is that it is not suggested that rejection of the supposed null hypothesis argues in favor of the non-proportional odds model as the alternative model. That model is not a viable alternative model, which leaves unanswered the fundamental question, what failure of the model does the Brant test reveal? Brant dwells on this question in his conclusion,

> As previously mentioned, assessment of the proportionality assumption can also be based on fitting the augmented models (2.1) [the non-proportional odds model], as in Hutchison (1985) and Ekholm and Palmgren (1989). Similarly, a more directed approach can be based on fitting (3.2) [Anderson's (1984) stereotype model]. The augmented model approach is attractive in that it provides a more standard theoretical framework for developing tests. One drawback, however, is that specialized algorithms must be developed to fit the augmented models. A more serious problem is inherent in the models themselves. For example, if one wishes to extend the use of model (2.1) beyond the values of **x**'s actually observed, the $\boldsymbol{\beta}_j$'s must be constrained to ensure monotonicity of the extrapolated γ_j's. Similar difficulties pertain to (3.2). Depending on the range of admissable values of **x**, this can lead to technical difficulties in fitting and the need for nonstandard likelihood theory to allow for the possibility of estimates falling on the boundary of the parameter space. *It may be best then to view (2.1) and (3.2) not as scientifically meaningful models, but as directional alternatives helpful in validating the simpler proportional odds model*. [Emphasis added.]

We conclude that the Brant test is useful for supporting or for casting doubt on the basic model. It does not seem to be useful for pointing toward what might appear superficially to be an alternative specification based on freeing the parameter vectors in γ_j.

We note, finally, the response of some analysts to the failure of the base model (the ordered choice model), say as evidenced by the Brant test, is to switch to the unordered multinomial logit model as an alternative. Williams (2006, p. 5) dismisses this approach because the alternative proliferates parameters and is difficult to interpret. In fact, switching to the multinomial logit model as an alternative to the ordered choice model, assuming that some ordered choice model was appropriate to begin with, substitutes a manifestly misspecified model for one that was merely suspect and, probably, in need of refinement. The multinomial logit model for unordered choices is applicable to a different situation entirely. It produces coefficients, but it would be arduous at best to translate them into something meaningful to describe the behavior of an ordered random variable, such as the outcome of an attitude survey. So, following Williams, we will eschew further consideration of the multinomial logit model for unordered choices in this review.

6.1.3 Generalized Ordered Logit Model (1)

Quednau (1988), Clogg and Shihadeh (1994), Fahrmeir and Tutz (1994), McCullagh and Nelder (1989) have proposed versions of the ordered choice models based essentially on the "non-proportional odds" form given above. Fu (1998) and Williams (2006) have recently provided working papers and a *Stata* program (`GOLogit` and `GOLogit2`) that implement and refine the model. Williams (2006) suggests that his development is an extension of Fu's so we focus on the latter. Motivated by the frequent rejection of the null hypothesis by Brant's (1990) test [see Williams (2006, p. 3)], a suggested alternative model derives from the core specification

$$\text{Prob}(y_i > j) = F(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)}, j = 0,1,...,J-1, \tag{6.12}$$

where, now, $\mathbf{x}_i$ does not contain a constant term. (Note that this is the form used by Brant to motivate his analysis.) The implication is

$$\begin{aligned}
\text{Prob}(y_i = 0|\mathbf{x}_i) &= 1 - F(\alpha_0 + \boldsymbol{\beta}_0'\mathbf{x}_i),\\
\text{Prob}(y_i = 1|\mathbf{x}_i) &= F(\alpha_0 + \boldsymbol{\beta}_0'\mathbf{x}_i) - F(\alpha_1 + \boldsymbol{\beta}_1'\mathbf{x}_i),\\
\text{Prob}(y_i = j|\mathbf{x}_i) &= F(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}'\mathbf{x}_i) - F(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i),\\
\text{Prob}(y_i = J|\mathbf{x}_i) &= F(\alpha_{J-1} + \boldsymbol{\beta}_{J-1}'\mathbf{x}_i).
\end{aligned}$$

We label this the "(1)" form of the generalized ordered choice model. We will examine two other forms, with (unfortunately) the same name. The "(1)" does not indicate first chronologically; that would be Terza's (1985) formulation. It is simply the first one presented in this review. This model is related to, but is not quite the same as the implied alternative in Brant's analysis. In fact, Brant's alternative model, which is equivalent to $\text{logit}(\gamma_{ij}) = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i$, treats each of the J+1 outcomes of y_i as a separate event – the probabilities vary completely independently and need not even sum to one or a number less than one. As he notes, it should not be viewed as a valid model as it stands. In the model suggested above, the ordering aspect of the observed variable is preserved somewhat, in that the formulation implies a connection between the events $y_i = j$ and $y_i = j$-1. On the other hand, with no constraints imposed on the parameters of the model, although the probabilities sum to one by construction, there is no assurance that they are positive. Brant anticipated this uncomfortable feature of the model in the conclusion related at the end of Section 6.1.2. Long and Freese (2006, p. 221) observe this as well, but note that "To ensure that the $\Pr(y=j|\mathbf{x})$ is between 0 and 1, the condition $(\tau_j - \boldsymbol{\beta}_j'\mathbf{x}) \geq (\tau_{j-1} - \boldsymbol{\beta}_{j-1}'\mathbf{x})$ must hold." (The inequality must actually be strong if the probabilities are to be nonzero as well.) Rewrite the restriction as $(\tau_j - \tau_{j-1}) > (\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1})'\mathbf{x}$. The only way to ensure that this is true for *every* possible configuration of $\mathbf{x}$ is to have $\tau_j > \tau_{j-1}$ and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j-1}$, which is where we began.

The problem of negative probabilities was raised much earlier. Williams (2006) invoking McCullagh and Nelder (1989, p. 155) observes

> "The usefulness of non-parallel regression models is limited to some extent by the fact that the lines must eventually intersect. Negative fitted values are then unavoidable for some values of **x**, though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the **x**-space, this flaw in the model need not be serious."
>
> This seems to be a fairly rare occurrence, and when it does occur there are often other problems with the model, e.g. the model is overly complicated and/or there are very small *N*s for some categories of the dependent variable. gologit2 will give a warning message whenever any in-sample predicted probabilities are negative. If it is just a few cases, it

> may not be worth worrying about, but if there are many cases you may wish to modify your model, data, or sample, or use a different statistical technique altogether.

The prescription relates to fitting the function to the data, but not to the underlying model. I.e., the "flaw" in the model is not that it sometimes produces negative fitted probabilities; it is that it does not impose the positivity of the fitted probabilities in the structure to begin with. In practical terms, as Williams (2006) suggests, the model is usually estimable, and the problem does not arise. If one begins the iterations with starting values obtained from the "constrained" ordered logit model, then at least at the starting values, one is assured that all probabilities are positive. As the iterations move away from the starting values, as any probability associated with an observed outcome moves toward zero, it will impose a large penalty on the log likelihood – in principle if a probability for an observation becomes negative, it exerts an infinite penalty. The practical upshot is that it seems reasonable that, in spite of its potential for internal inconsistency, this model is likely to be estimable. Table 6.2 shows the results for our ordered choice example. (Williams (2006) has published a *Stata* program (`GOLogit2`) for this purpose. We used the `MAXIMIZE` command in *NLOGIT*.) The estimates in Table 6.2 have been reordered so that coefficients associated with specific independent variables are grouped contiguously, rather than coefficients associated with specific outcomes. Inspection of the sets of estimates certainly suggests that the coefficients differ substantially across j. A likelihood ratio test would be based on

$$\chi^2[15] = 2(-5713.579 - (-5752.985)) = 78.812.$$

The 95% critical value from the table is 24.996. Thus, the hypothesis of the restricted model is decisively rejected.

A peculiarity of this "generalization" of the ordered logit model is that it does not appear to define a random variable. The specification states that "If $y_i = j$, then the probability that y_i equals j is as follows: "In spite of its appearance, the model does not state that the probability that a well defined random variable is equal to the given value is equal to the function. There is no underlying continuous variable that can be structured so as to produce the observed outcome. The latent regression approach is not available to motivate the outcome variable; "$y^* = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x} + \varepsilon$ then $y^* = j$ under some condition," since in order to generate y^*, one would need to know the appropriate j in advance. Consider, for example, that it is not possible to simulate the values of the random variable, y, defined in the probability statement. In order to assign a probability to the outcome we would first have to know what the outcome is. No data generating process produces the random variable described in the probability statement. This model, *as stated*, has the uncomfortable feature that it does not define what the "random variable "y" is; it defines y in terms of itself. Ultimately, the problem is the ordered nature of the observed response. The ordering is incompatible with that much free parameter variation in the statement of the probabilities. If a model of an ordered random variable is to be complete and internally consistent, then ultimately the observed response must be derived as a classification of a set of underlying events. The early writers on this model, Aitchison, McCullagh, Snell, etc., returned repeatedly to the theme of the underlying continuous variable for this reason.

Table 6.2 Estimated Ordered Logit and Generalized Ordered Logit (1)

Ordered Probability Model	
Underlying probabilities based on Logit	
Dependent variable	HEALTH
Log likelihood function	-5749.187
Restricted log likelihood	-5875.096
Chi squared	251.8798
Degrees of freedom	5
Prob[ChiSqd > value] =	.0000000

Variable	Coefficient	Standard Error	b/St.Er	P[|Z|>z]
Index function for probability				
Constant	3.5179	.2038	17.260	.0000
AGE	-.0321	.0029	-11.178	.0000
EDUC	.0645	.0125	5.174	.0000
INCOME	.4263	.1865	2.286	.0223
MARRIED	-.0645	.0746	-.865	.3868
KIDS	.1148	.0669	1.717	.0861
Threshold parameters for index				
Mu(1)	2.1213	.0371	57.249	.0000
Mu(2)	4.4346	.0390	113.645	.0000
Mu(3)	5.3771	.0520	103.421	.0000

User Defined Optimization		Generalized Ordered Logit
Maximum Likelihood Estimates		Logit Model (1)
Log likelihood function	-5713.579	

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Ordered Logit Estimates
Constant	2.69537	.606874	4.441	.0000	α = 3.51646
	1.04676	.251309	4.165	.0000	$\alpha-\mu_1$ = 1.39503
	-.67133	.253798	-2.645	.0082	$\alpha-\mu_2$ = -.91697
	-1.09368	.368911	-2.965	.0030	$\alpha-\mu_3$ =-1.86024
AGE	-.04080	.007651	-5.332	.0000	AGE
	-.02925	.003426	-8.538	.0000	-0.03213
	-.03261	.003758	-8.677	.0000	
	-.02427	.004968	-4.885	.0000	
EDUC	.12009	.038709	3.102	.0019	EDUC
	.07635	.015527	4.917	.0000	0.06467
	.06222	.015730	3.956	.0001	
	-.00252	.023385	-.108	.9141	
INCOME	1.98158	.452708	4.377	.0000	INCOME
	.51201	.214586	2.386	.0170	0.42434
	.18838	.233611	.806	.4200	
	-.11631	.285676	-.407	.6839	
MARRIED	.05870	.171015	.343	.7314	MARRIED
	-.02514	.086290	-.291	.7708	-.06451
	-.15166	.096590	-1.570	.1164	
	-.07179	.129624	-.554	.5797	
KIDS	.34731	.184095	1.887	.0592	KIDS
	.21913	.081866	2.677	.0074	0.11452
	.01939	.088280	.220	.8261	
	-.11322	.121602	-.931	.3518	

In our application, we began the computations by collapsing several categories of the dependent variable, for example, combining categories 0,1,2 into the observed "0." Likewise, Boes and Winkelmann (2006a) combined the lowest three categories of their observed satisfaction measure. The implication of the generalized ordered probability model (1) would be either that in the collapsed model, the coefficient vector associated with the zero outcome is an ambiguous mixture of the original three coefficient vectors, or in the original model, the lowest three categories have the same coefficient vector – that would legitimize the aggregation of the three cells. It is a matter of interpretation. The implication, however, is that the population parameters $(\alpha_j,\boldsymbol{\beta}_j)$ exists only as a function of the way that the analyst codes the dependent variable. More to the point, the model parameters, e.g., the data generating mechanism, cannot consistently exist apart from the observed data themselves. This returns to the characteristic that it is not possible to simulate a well defined random variable that obeys the probability laws defined for this model. This might seem to be true in the base case model, since the "cut points" are identified with the outcomes. However, it is not the case there, since μ_j exists (in theory) as an unknown location on the real line, independently of the random variable that drives the model, $y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$. There is no counterpart to y^* in the Generalized Ordered Logit Model (1).

All this said, it remains true that the "parameters" of the model *can* be computed, as we have done in Table 6.2. The least favorable view is that this is just curve fitting. However, if so, and if the ordered logit model (same $\boldsymbol{\beta}$) really is appropriate, then one should replicate, at least approximately the original "constrained" model. To some degrees, as evident in Table 6.2, that is what occurs; this could be viewed as a (numerically) inefficient estimator of the original model. But, in the same spirit as the Brant test, the same question emerges. To the extent that this procedure does not mimic the original model – the separate parameter vectors really do differ, as ours do in Table 6.2 – then what has it found? Since the model, such as it is, is not a valid probability model, the same loose end emerges. It must be picking up *some* failure of the original model. One might guess that Brant's speculations about a set of explanations for rejection of the null hypothesis by his test would be helpful here as well.

We have labeled the model discussed here the "Generalized Ordered Choice Model (1)." Forms "(2)" and "(3)" are discussed below. The preceding is an orthodox interpretation of the model specification. Later, in Section 7.3, we will find that with a straightforward reinterpretation of what is ultimately the same model structure, an internally consistent specification of a random variable does emerge. Since the models are only superficially different, we will label the threshold models in Section 7.3 the "(2)" forms of the Generalized Ordered Choice Model."

6.2 Model Implications for Partial Effects

Superficially, it seems that the ordered choice model is using a single index function, $\boldsymbol{\beta}'\mathbf{x}_i$, to describe the determination of $J+1$ outcomes, $y = j$. Even though in fact, there is only a single outcome, $y_i = T(y_i^*)$, it remains interesting to examine the particular values that y_i attains. For example, the analyst is often specifically in the highest or lowest cell. Brewer et al. (2008) were interested in the top several cells in the distribution. As we noted earlier, since there is no single natural conditional mean function, the typical analysis describes the probabilities individually with the partial effects described in Chapter 5. Because the model is a 'single index' specification – there is only one $\boldsymbol{\beta}'\mathbf{x}_i$ in the model – a large number of constraints are imposed on the partial effects..

6.2.1 The Single Crossing Feature of the Ordered Choice Model

The partial effects shown in the preceding examples vary with the data and the parameters. Since the probabilities must sum to one, the partial effects for each variable must sum to zero across the probabilities. It can also be shown that for the probit and logit models, this set of partial derivatives will change sign exactly once in the sequence from 0 to *J*, a property that Boes and Winkelmann (2006b) label the *single crossing* characteristic. [Crawford, Pollak and Vella (1988) explore this feature of the model at length.] For a positive coefficient, β_k, the signs moving from 0 to *J* will begin with negative and switch once to positive at some point in the sequence. The following Table 6.3 is extracted from Table 4 in Boes and Winkelmann (2006b, page 22). (The "0-2" bracket is obtained by grouping the relatively low number of observations with the three lowest values in the original data.) Partial effects are shown with estimated standard errors in parentheses.

Table 6.3 Boes and Winkelmann Estimated Partial Effects

Response	0-2	3	4	5	6	7	8	9	10
Men									
OProbit	-0.016	-0.014	-0.016	-0.037	-0.020	0.003	0.059	0.027	0.014
	(0.003)	(0.001)	(0.001)	(0.003)	(0.009)	(0.003)	(0.009)	(0.005)	(0.005)
GOProbit	-0.020	-0.022	-0.014	-0.027	-0.037	-0.005	0.088	0.039	-0.002
	(0.007)	(0.006)	(0.004)	(0.005)	(0.006)	(0.007)	(0.033)	(0.109)	(0.089)
Women									
OProbit	-0.004	-0.005	-0.005	-0.016	-0.008	-0.003	0.020	0.012	0.008
	(0.002)	(0.001)	(0.001)	(0.005)	(0.012)	(0.003)	(0.011)	(0.004)	(0.006)
GOProbit	-0.009	0.005	-0.011	-0.036	-0.040	0.038	0.064	-0.008	-0.003
	(0.008)	(0.016)	(0.020)	(0.015)	(0.013)	(0.029)	(0.116)	(0.125)	(0.027)

The same effect can be seen in Table 5.2 for our application.

The "`GOProbit`" results – a probit version of Williams's (2006) `GOLogit` approach – show the effect of relaxing the single crossing restriction. However, for men, the model seems to be preserving the restriction on its own – the second crossing at $y = 10$, produces a marginal effect that differs only trivially from zero, with a "*z*-value" of only 0.022. For women, however, one is in the uncomfortable position of now explaining four crossings which make the model seem somewhat unstable. None of the estimated effects are statistically significant, in contrast to the ordered probit model, and in fact, two of the crossings rest on what looks like a maverick finite sample outcome at y=3. One the other hand, the results that remain force the analyst into a counterintuitive position of arguing that higher incomes are associated with lowered probabilities of reporting a high subjective well being – perhaps a widespread *Richard Cory* effect. The authors' description of the results (from their pages 12 and 13) suggests the appeal of a less sharp statement about specific outcomes; the right tail result is suggested to reflect a zero effect, which of course removes the remaining extra crossing:

> Table 4 summarizes the marginal probability effects of income by gender. Consider, for example, the results for men and take the ceteris paribus effect of increasing logarithmic household income by a small amount on the probability of responding a SWB level of "8". Table 4 shows a value of 0.059 for the standard model. This means that the probability of a response of "8" increases by 0.059 percentage points if we increase logarithmic income by 0.01, which corresponds approximately to a one-percent increase in level income. A doubling of income, i.e., a change in logarithmic income by 0.693, increases the probability of response "8" by about 0.059×0.693×100, or about 4.09 percentage points, ceteris paribus.
>
> Comparing the MPE's among the three different models and over all possible outcomes, we obtain the following main results. For men all models suggest that more income significantly reduces the probability of low SWB (0-5), and significantly increases the probability of response "8". For high SWB responses (9-10), the standard

> model predicts a strong positive relationship between income and SWB, whereas the generalized model and also the binary models do not find a significant effect. Since the restricted OProbit is clearly rejected, we conclude that income has no effect on positive well-being. *Our preferred specification supports the asymmetry hypothesis for men: higher income decreases the probability of negative well-being (low SWB), but it does not affect the probability of positive well-being (high SWB).* [Emphasis added.] For women the relationship between income and SWB is relatively weak. While the standard model finds small but significant effects for low and high SWB responses, *the generalized model predicts a significant negative effect only on the probability of responses "5" and "6"*. [Emphasis added.] The gender difference might be explained by social norms that assign the role of primary income earner to men and therefore make income a relatively more important determinant of male well-being (see also Lalive and Stutzer 2004).

Figure 6.1 shows graphically the values in Table 6.3. The ordered probit and generalized ordered probit models do not seem to be giving different accounts. The latter does seem to be exaggerating the outcome at choice 8, or perhaps suggesting a significant spike associated with that outcome, that needs some explanation. The force of the model extension seems to be to produce a much more pronounced effect in the middle of the distribution. The fact that the heightened impact is negative for $y = 6$ and positive for $y = 8$, followed for both genders by a sharp return to zero at $y = 9$, seems a bit counterintuitive.

The shortcoming of the ordered choice model that produces the single crossing result is the linearity of the single index formulation. One can achieve the same result as above without resort to the generalized model simply by building the desired curvature into the index function itself. In the figure below, we have re-estimated our original model based not on using "Health" coded 0 to 4, but the original Health Satisfaction variable, coded 0 to 10, the same as in Boes and Winkelmann's study. (They are subsets of the same data base.) Income is included in linear, squared and cubed form, so that the marginal effect of income on any outcome is

$$\delta_{INCOME}(j) = [f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x})] \times (\beta_{INCOME} + 2\beta_{INCOME\text{-}SQ}\, INCOME + 3\beta_{INCOME\text{-}CUBE}\, INCOME^2).$$

We have evaluated this at the means of all the variables in the model. The results are shown in Figure 6.2 along with the results from the original model. While the effects still only cross zero once, the formulation does not force this – we will accept the data's word for it that the partial effect of income does indeed (at least seem to) start negative and become positive, conforming to intuition that greater income is broadly associated with greater health satisfaction. It is interesting as well that the linear index model produces essentially the same results.

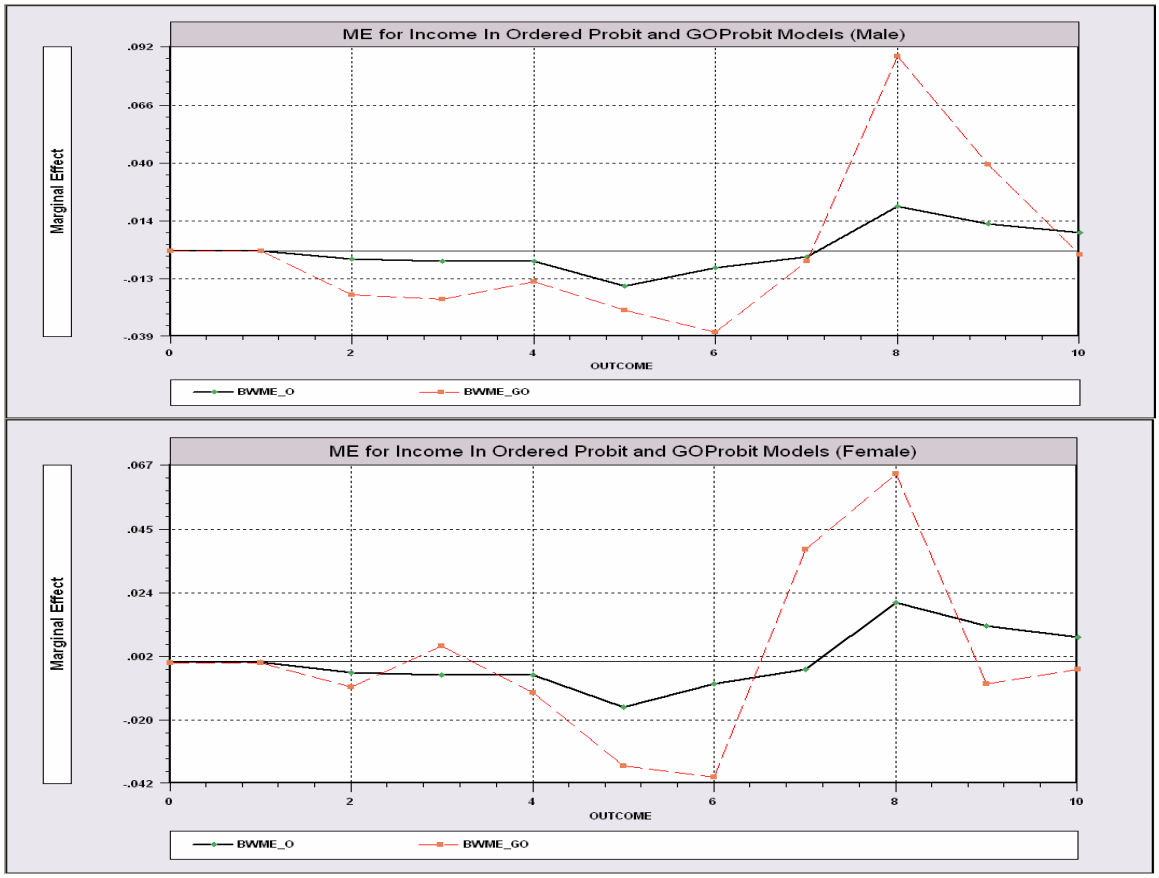


Figure 6.1 Estimated Partial Effects in Boes and Winkelmann (2006b) Models

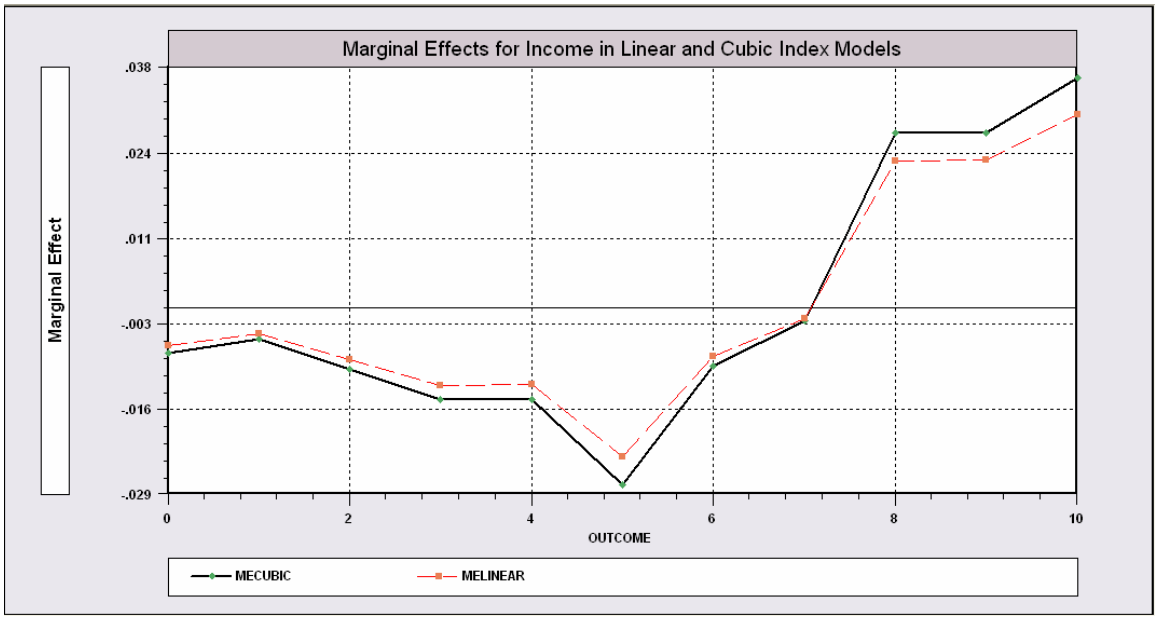


Figure 6.2 Estimated Partial Effects for Linear and Nonlinear Index Functions

6.2.2 Choice Invariant Ratios of Partial Effects

Boes and Winkelmann (2006a) note that for any two continuous covariates, x_{ik} and x_{il}

$$\frac{\partial \text{Prob}[y_i = j \mid \mathbf{x}_i]/\partial x_{i,k}}{\partial \text{Prob}[y_i = j \mid \mathbf{x}_i]/\partial x_{i,l}} = \frac{\beta_k}{\beta_l},$$

which is independent of the outcomes. This is a feature of the assumed underlying utility function, the same as in any regression model. Any single index function model that is of the form

$$\text{Prob}(y_i = j|\mathbf{x}) = G_j(\boldsymbol{\beta}'\mathbf{x}_i),$$

will have this feature; it is a consequence of the chain rule of the calculus. It is unclear what the behavioral implications will be; that would be specific to the application. Boes and Winkelmann (2006a) develop this theme in some detail. In their application to subjective well being,

$$SWB = \alpha + \beta_{INCOME} INCOME + \beta_{UNEMPLOYMENT} UNEMPLOYED + ... + \varepsilon,$$

the authors are interested in the notion of "compensating variation." For their purpose, "what is the income increase required to offset the negative well-being effect of unemployment?" (They finesse the binary nature of the unemployment variable by considering the issue from the point of view of the population unemployment *rate*.) By equating the total differential of Prob($y = j|\mathbf{x}$) to zero, they find that the interesting "tradeoff ratio" is the negative of the ratio of the partial effects, as shown above. The implication of the standard model is that the tradeoff ratios are the same for all outcomes.

In the semiparametric models developed in Chapter 12, in which it is not possible to compute the CDF or the density – the semiparametric aspect of the model is to dispense with the assumption of a specific density – ratios of coefficients become important outputs of the estimation process. Stewart (2003, 2005) develops this idea at some length.

The common feature of this and the extensions preceding it are that the functional form is built around the outcomes. The single index models considered thus far do not provide sufficient curvature to accommodate what Anderson (1984) called the "dimensionality" of the problem. The greater fit achieved by the expanded model may have less to do with describing the underlying data generating process than with matching the fitted function to the pattern in the observed data. The modifications of the ordered choice model described in the next several chapters also achieve some of this increased "fit" but do so within the structure of the original behavioral model.

6.3 Methodological Issues

The various generalizations of the model suggested above do deal with the problems of parallel regressions and single crossing, but potentially create new ones. The heterogeneity in the parameter vector is an artifact of the coding of the dependent variable, not a manifestation of underlying heterogeneity in the dependent variable induced by behavioral differences. It is unclear what it means for the marginal utility parameters to be structured in this way. To put a better face on it, we might better interpret this as a semiparametric approach to modeling what is apparently underlying heterogeneity, however, again, it is not clear why this should be manifest in parameter variation across the outcomes instead of across the individuals in the sample. One would assume that the failure of the Brant test to support the model with parameter homogeneity

is, indeed, signalling some failure of the model. But, it is unclear what that failure is. The more difficult problem of this generalization of the model is that the probabilities in this model need not be positive, and there is no parametric restriction (other than the restrictive model one we started with) that could achieve this. The restrictions would have to be functions of the data. (The problem is noted by Williams (2006), but dismissed as a minor issue. Boes and Winkelmann suggest that the problem could be handled through a "nonlinear specification.")

One might still argue that there are differences across the individuals at the "low" end vs. the "high" end of the distribution. The excerpt from Boes and Winkelmann above would suggest this. In fact, the single crossing aspect of the model accommodates this feature. Still, something more akin to a latent class structure would seem to apply under this interpretation. In such a setting, one is likely to find that the high outcomes are more likely for some classes then others. The advantage of this approach would be that the class structure can be assumed to be exogenous. One is not forced to make the model structure endogenous to the observed outcomes.

6.4 Specification Tests for Ordered Choice Models

The ordered probit model is a conventional model by the standards of maximum likelihood estimation. Under the assumptions that the model is correctly specified and the data on y_i and $\mathbf{x}_i$ are "well behaved," [see Greene (2008a, chapter 4)], the familiar asymptotics and testing procedures used in Section 5.6 apply. That is, we can use the familiar apparatus, namely Wald, Lagrange multiplier and likelihood ratio procedures to test against null hypotheses that are nested within the essential parametric model,

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) > 0, j = 0,1,...,J. \quad (6.13)$$

Since the asymptotic theory relies on central limit theorems, and not on the specific distribution of ε_i, the same devices will apply for logit and probit models. Procedures for "exact" inference based specifically on the distribution assumed [see, e.g., Mehta and Patel (1995)] have not been developed for ordered choice models.

In this section, we consider "specification tests." That is, tests against the null specification of the model, for which often there is no clearly defined alternative. For example, a test of the appropriateness of the assumption that ε_i is normally distributed is considered against the alternative that it is not. Specification tests for the ordered choice model have been obtained essentially for two issues, functional form and distribution. The functional form question relates to the assumption about the basic model specification,

$$\text{Prob}(y_i > j \mid \mathbf{x}_i) = F(\boldsymbol{\beta}'\mathbf{x}_i - \mu_j), j = 0,...,J\text{-}1. \quad (6.14)$$

The linearity of the index function is the main issue, though it will be clear shortly that, because the alternative hypothesis is not clearly stated, a test against this null might pick up a variety of other failures of the model assumption. The distributional tests are specifically directed to the question of whether normality (or logisticality) is appropriate. Once again, the alternative hypothesis is unclear. For example, it seems reasonable to suggest that a test against normality might be picking up the influence of an omitted variable – perhaps one with a skewed distribution. Recognizing the essential ambiguity of the nature of these tests, we can nonetheless usefully divide them into these two broad groupings.

We note in passing, a third type of specification test that has been considered. Section 9.3 discusses a counterpart to the Hausman (1978) test for random vs. fixed effects in a panel data model. Since the test is considered in detail there, we will not reconsider it in this section.

6.4.1 Model Specifications – Missing Variables and Heteroscedasticity

A number of studies have considered the null specification of the ordered choice models against specific alternatives. These tests involve three particular features of the model, missing variables, heteroscedasticity and the distribution of ε_i. Murphy (1994, 1996), for example, examines the ordered logit model as a special case of the more general model

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}'\mathbf{z}_i + \sigma_i \varepsilon_i,$$
$$y_i = j \text{ if } \mu_{j-1} < y_i^* \leq \mu_j,$$

where

(1) $\mathbf{z}_i$ is a set of omitted variables that are believed to be appropriate to be in the model;

(2) $\sigma_i^2 = (\pi^2/3)[\exp(\boldsymbol{\alpha}'\mathbf{h}_i)]^2$;

(3) $F(\varepsilon_i) = [1 + \exp(-\varepsilon_i)]^{-\delta}$. (This is an asymmetric distribution.)

Murphy's extended ordered logit model encompasses the familiar ordered logit model; the null hypothesis of the restricted model would be $\boldsymbol{\gamma} = \mathbf{0}$, $\boldsymbol{\alpha} = \mathbf{0}$, $\delta = 1$. In principle, the alternative model can be fit by full information maximum likelihood. If so, then the tests of the three specifications can be done one at a time or jointly, using Wald or Likelihood ratio tests. Murphy proposes Lagrange multiplier tests for the three hypotheses that involve only estimating the restricted, basic model. We will consider the missing variables and heteroscedasticity tests here, and return to the distribution in the next section.

For the moment, we revert to the simpler distribution with $\delta = 1$, and examine the *LM* test for missing variables and heteroscedasticity. Without the special consideration of the shape of the distribution (δ), the testing procedures are the same for the probit and logit models, so they are given generically below. In this context, it is worth noting, since $\mathbf{z}_i$ is observed, not much is gained by using an *LM* test for missing variables; one can just as easily fit the full model and use the *LM* or Wald test of the null hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$. The test for heteroscedasticity is likewise straightforward if one is able to fit the full model with this form of heteroscedasticity. [The *LM* tests proposed by Murphy (1994, 1996) and Weiss (1997) actually apply to any form of heteroscedsticity such that $\sigma_i^2 = \sigma_0^2 w(\boldsymbol{\gamma}, \mathbf{h}_i)$ such that $w(\mathbf{0},\mathbf{h}_i) = 1$. [See Breusch and Pagan (1979).] Harvey's (1976) model has been the form usually used in the received applications.

Consider, first, an *LM* test for missing variables. The log likelihood function is

$$\log L = \sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \log\left[F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i)\right]. \quad (6.15)$$

The *LM* test is carried out by estimating the model under the null hypothesis that $\boldsymbol{\gamma} = 0$, then obtaining the statistic,

$$LM = \left(Est.\partial \log L \Big/ \partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\mu}\\ \boldsymbol{\gamma}\end{pmatrix}\right)'_{|\boldsymbol{\gamma}=\mathbf{0}} \left\{Est.Var\left[\partial \log L \Big/ \partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\mu}\\ \boldsymbol{\gamma}\end{pmatrix}\right]\right\}^{-1} \left(Est.\partial \log L \Big/ \partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\mu}\\ \boldsymbol{\gamma}\end{pmatrix}\right)_{|\boldsymbol{\gamma}=\mathbf{0}}. \quad (6.16)$$

(We have reversed the usual order of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ for convenience.) The test statistic is used to test the hypothesis that the gradient is zero at the restricted parameter vector. When the restricted

model is fit by maximum likelihood, the derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ evaluated at the MLEs are numerically zero, so the sample estimator of the statistic is

$$LM = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \dfrac{\partial \log L}{\partial \boldsymbol{\gamma}} \end{pmatrix}'_{|\boldsymbol{\gamma}=\mathbf{0}} \left\{ Est.Var\left[\partial \log L \Big/ \partial \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\mu} \\ \boldsymbol{\gamma} \end{pmatrix}_{|\boldsymbol{\gamma}=0} \right] \right\}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \dfrac{\partial \log L}{\partial \boldsymbol{\gamma}} \end{pmatrix}_{|\boldsymbol{\gamma}=\mathbf{0}} . \tag{6.17}$$

The practical application of the test requires computation of the derivatives of the log likelihood with respect to $\boldsymbol{\gamma}$, evaluated at $\boldsymbol{\gamma} = 0$, and an estimator of the asymptotic covariance matrix, which we consider below. For the derivatives,

$$\begin{aligned} \left(\frac{\partial \log L}{\partial \boldsymbol{\gamma}} \right)_{|\boldsymbol{\gamma}=\mathbf{0}} &= \left\{ \sum_{i=1}^{N} \sum_{j=0}^{J} m_{ij} \left[\frac{f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i) - f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i)}{F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}'\mathbf{z}_i)} \right] (-\mathbf{z}_i) \right\}_{|\boldsymbol{\gamma}=\mathbf{0}} \\ &= \sum_{i=1}^{N} \sum_{j=0}^{J} m_{ij} \left[\frac{f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)}{F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)} \right] (-\mathbf{z}_i). \end{aligned} \tag{6.18}$$

It remains to obtain the appropriate asymptotic covariance matrix. A convenient estimator is the sum of the outer products of the individual gradients, $\mathbf{H}^{-1} = [\Sigma_i \, \mathbf{g}_i \mathbf{g}_i']^{-1}$. (However, Davidson and MacKinnon (1983, 1984), MacKinnon (1992), Godfrey (1988), and Weiss (1997) present evidence that the finite sample properties of the *LM* statistic are inferior to those when it is based on the second derivatives matrix or, when possible, the expected second derivatives matrix.) Precise expressions for the second derivatives matrix appear in various places, including McElvey and Zavoina (1975), Maddala (1983) and in (5.16) in Section 5.9.5. Write the second derivatives matrix in the partitioned form

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\beta\beta'} & \mathbf{H}_{\beta\mu'} & \mathbf{H}_{\beta\gamma'} \\ \mathbf{H}_{\mu\beta'} & \mathbf{H}_{\mu\mu'} & \mathbf{H}_{\mu\gamma'} \\ \mathbf{H}_{\gamma\beta'} & \mathbf{H}_{\gamma\mu'} & \mathbf{H}_{\gamma\gamma'} \end{bmatrix}.$$

Then, from the form of the first derivatives, it can be seen that the *LM* statistic equals the first derivatives vector times the lower right submatrix of $\mathbf{H}^{-1}$. Collecting terms and using the partitioned inverse form [Greene (2008a, result A-74)], this will be

$$LM = \left\{ \left(\frac{\partial \log L}{\partial \boldsymbol{\gamma}} \right)_{|\boldsymbol{\gamma}=\mathbf{0}} \right\}' \left[\mathbf{H}_{\gamma\gamma'} - \begin{pmatrix} \mathbf{H}_{\gamma\beta'} & \mathbf{H}_{\gamma\mu'} \end{pmatrix} \begin{bmatrix} \mathbf{H}_{\beta\beta'} & \mathbf{H}_{\beta\mu'} \\ \mathbf{H}_{\mu\beta'} & \mathbf{H}_{\mu\mu'} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{H}_{\beta\gamma'} \\ \mathbf{H}_{\mu\gamma'} \end{pmatrix} \right]^{-1} \left\{ \left(\frac{\partial \log L}{\partial \boldsymbol{\gamma}} \right)_{|\boldsymbol{\gamma}=\mathbf{0}} \right\}. \tag{6.19}$$

Weiss (1997) notes an interesting interpretation of the *LM* test for omitted variables. The gradient, $(\partial \log L/\partial \boldsymbol{\gamma})|\boldsymbol{\gamma}=0$ given earlier can be written

$$\left(\frac{\partial \log L}{\partial \gamma}\right)_{|\gamma=0} = \sum_{i=1}^{N} \sum_{j=1}^{J} m_{ij}\left[\frac{f(\mu_{j-1}-\beta'\mathbf{x}_i)-f(\mu_j-\beta'\mathbf{x}_i)}{F(\mu_j-\beta'\mathbf{x}_i)-F(\mu_{j-1}-\beta'\mathbf{x}_i)}\right](\mathbf{z}_i)$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{J} m_{ij} E[\varepsilon_i \mid \mathbf{x}_i, y_i = j](\mathbf{z}_i) \qquad (6.20)$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{J} m_{ij} GR(1)_{ij}\mathbf{z}_i.$$

That is, the test is based on the covariance between the (unobserved) disturbance and the omitted variables. This is precisely the approach used in the linear regression model, where ε_i is estimated directly with the residual, e_i. In this case, the estimator is a "*generalized residual*." [See Chesher and Irish (1987) and Gourieroux et al. (1987).]

An *LM* test for heteroscedasticity is essentially the same, save for the considerably more complicated first and second derivatives. The model with heteroscedasticity (and no missing variables) has

$$\log L = \sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \log\left[F\left(\frac{\mu_j-\beta'\mathbf{x}_i}{\sigma_i}\right)-F\left(\frac{\mu_{j-1}-\beta'\mathbf{x}_i}{\sigma_i}\right)\right]. \qquad (6.21)$$
$$= \sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \log\left[F_{i,j}-F_{i,j-1}\right]$$

The first derivative vector is

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} \sum_{j=0}^{J} \left[m_{ij}\frac{f_{i,j}\times\left(\frac{\mu_j-\beta'\mathbf{x}_i}{\sigma_i}\right)-f_{i,j-1}\times\left(\frac{\mu_{j-1}-\beta'\mathbf{x}_i}{\sigma_i}\right)}{F_{i,j}-F_{i,j-1}}\right](-\mathbf{h}_i). \qquad (6.22)$$

The remaining computations are analogous to those done for the missing variables test. Note that under the null hypothesis, $\sigma_i = 1$, which considerably simplifies computing (albeit not deriving) the first and second derivatives.

In many cases, test statistics such as the *LM* statistic are computable using "artificial regressions." For many of the common applications, we may write the LM statistic in the form

$$LM = \left(\sum_{i=1}^{n} w_i(\theta)\mathbf{g}_i(\theta)\right)'\left(\sum_{i=1}^{n} \mathbf{g}_i(\theta)\mathbf{g}'(\theta)_i\right)^{-1}\left(\sum_{i=1}^{n} w_i(\theta)\mathbf{g}_i(\theta)\right), \qquad (6.23)$$

where $\boldsymbol{\theta}$ is the full parameter vector being analyzed. In this case, the LM statistic is equal to the explained sum of squares in the regression of the variable $w_i(\boldsymbol{\theta})$ on the "pseudo-regressors," $\mathbf{g}_i(\boldsymbol{\theta})$. [See MacKinnon (1992) and Orme (1990).] Consider the narrower case in which $\mathbf{g}_i(\boldsymbol{\theta})$ is the full gradient, $w_i(\boldsymbol{\theta}) = 1$, and the outer product of gradients (OPG) estimator of the covariance matrix is used to complete the statistic. Then, the "dependent variable" in this regression is 1 for all *i*, the total uncentered sum of squares is *n*, and $LM = nR^2$ in the artificial regression. [See Davidson and MacKinnon (1984).] With the extensive matrix manipulation routines in contemporary software such as *Stata* (Mata), *SAS* (Proc MATRIX), *NLOGIT* (Matrix), *Gauss* and *Matlab*, the appeal of the artificial regression interpretation is now largely confined to the analytics that precede computation.

6.4.2 Testing Against the Logistic and Normal Distributions

Murphy (1994, 1996) proposes an alternative distribution for ε in the ordered logit model,

$$F(\varepsilon) = \frac{1}{[1+\exp(-\varepsilon)]^{\delta}}.$$

This distribution of ε_i is asymmetric; called a Burr type II distribution. This has been labeled the "scobit model" (skewed logit) elsewhere and has been suggested as an alternative to the normal and logistic distributions for binary choice models. [See Murphy (1994), Smith (1989), Lechner (1991), Nagler (1994) and *Stata* (2008) or Econometric Software (2007).] The density is

$$f(\varepsilon) = \left(\frac{\exp(-\varepsilon)}{1+\exp(-\varepsilon)}\right)\frac{\delta}{[1+\exp(-\varepsilon)]^{\delta}}.$$

For δ = 1, the model reverts to the familiar logit form. Since this is fully parameterized, the alternative model can be fit directly and a Wald or likelihood ratio test can be used to test the null hypothesis that δ = 1. Murphy proposes a Lagrange multiplier test that is based entirely on computations from the ordered logit model (δ = 1).

The scobit model has not been widely used in the ordered choice literature; tests about the distribution generally revolve around alternatives to the normal. Tests of the normality assumption build on the approach developed by Bera, Jarque and Lee (1984) for limited dependent variable models. A parametric alternative to the normal distribution is the Pearson family of distributions,

$$f(\varepsilon) = \frac{\exp[q(\varepsilon)]}{\int_{-\infty}^{\infty}\exp[q(t)]dt}, \text{ where } q(t) = \int \frac{c_1 - t}{c_0 - c_1 t + c_2 t^2}dt.$$

The relationship between the moments of the random variable and the three constants is

$c_0 = (4\tau_4 - 3\tau_3^2) / (10\tau_4 - 12\tau_3^2 - 18)$,
$c_1 = \tau_3 (\tau_4 + 3) / (10\tau_4 - 12\tau_3^2 - 18)$,
$c_2 = (2\tau_4 - 3\tau_3^2 - 6) / (10\tau_4 - 12\tau_3^2 - 18)$.

[See Weiss (1997).] (We are avoiding a potentially confusing conflict in notation by using τ rather than the conventional μ to denote the moments of the distribution.) For the standard normal distribution, $\tau_3 = 0$ and $\tau_4 = 3$. It follows that $c_0 = 1$, $c_1 = 0$ and $c_2 = 0$. (It also follows that the functional form is that of the standard normal.) Bera et al. (1984) developed an *LM* test for this restriction for the censored regression model. The corresponding result for the ordered probit model is given in Johnson (1996), Glewwe (1997) and Weiss (1997).

The test is based on the generalized residuals. For the normal distribution, we are tesing against the hypothesis that the third and fourth moments of ε are $\tau_3 = 0$ and $\tau_4 = 3$. As before, we cannot observe ε, so the test is based on the generalized residuals,

$$
\begin{aligned}
E[\varepsilon_i^3 \mid y_i = j, \mathbf{x}_i] \quad &= GR(3)_{ij} = \frac{\left\{\begin{array}{l}[2 + (\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)^2]\phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - \\ [2 + (\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)^2]\phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)\end{array}\right\}}{\Phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)}, \\
E[\varepsilon^4 - 3 \mid y_i = j, \mathbf{x}] &= GR(4)_{ij} = \frac{\left\{\begin{array}{l}[3 + (\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)^2](\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)\phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - \\ [3 + (\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)^2](\mu_j - \boldsymbol{\beta}'\mathbf{x})\phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)\end{array}\right\}}{\Phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)}.
\end{aligned}
\tag{6.24}
$$

The full derivative vector including c_1 and c_2 evaluated at $c_1 = c_2 = 0$ is

$$
\partial \log L / \partial \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\mu} \\ c_1 \\ c_2 \end{pmatrix} = \sum_{i=1}^{N} \left[\sum_{j=0}^{J} m_{ij} \begin{pmatrix} GR(1)_{ij}\mathbf{x}_i \\ GR(1)_{ij}\mathbf{a}_j \\ GR(3)_{ij} \\ GR(4)_{ij} \end{pmatrix} \right] = \sum_{i=1}^{N} \mathbf{g}_i. \tag{6.25}
$$

where $\mathbf{a}_j$ is a $(J\text{-}1)\times 1$ vector that has a 1 in position j and a -1 in position j -1 save for $j = 1$, when the j-1 position is absent. To complete the computation of the test statistic, an estimator of the covariance matrix of the gradient is needed. Notwithstanding its less than ideal finite sample properties, the usual choice is the outer products matrix,

$$
\mathbf{V} = \sum_{i=1}^{N} \mathbf{g}_i \mathbf{g}_i'.
$$

Using the maximum likelihood estimates from the ordered probit model, the first two parts of the derivative vector will be numerically zero. This, the final result for the LM statistic is

$$
LM = \left[\sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \begin{pmatrix} GR(3)_{ij} \\ GR(4)_{ij} \end{pmatrix} \right]' \left[\mathbf{V}_{c_0, c_1}^{-1} \right] \left[\sum_{i=1}^{n} \sum_{j=0}^{J} m_{ij} \begin{pmatrix} GR(3)_{ij} \\ GR(4)_{ij} \end{pmatrix} \right], \tag{6.26}
$$

where $\mathbf{V}_{c_1, c_2}^{-1}$ denotes the southeast 2×2 submatrix of $\mathbf{V}^{-1}$.

Glewwe (1997) discusses other methods of testing for normality (actually symmetry, $\tau_3 = 0$, and mesokurtosis, $\tau_4 = 3$) without a full parameterization of the alternative hypothesis, by using conditional moment tests. Newey (1985), Tauchen (1985) and Pagan and Vella (1989) provide details. As Glewwe shows, the *LM* test is essentially the same test. The use of the generalized residuals above suggests why this should be expected. Even though the *LM* test is structured around the Pearson alternative, in the end, it is a test of the values of the third and moments. The use of conditional moment tests is also pursued by Mora and Moro-Egido (2008). For J of the J+1 outcomes (because one is redundant), the model implies a set of moment conditions,

$$
E[m_{ij} - P_{ij}(\boldsymbol{\theta})] = 0,
$$

based on the additional assumptions of the model that produces the precise form of the probabilities. The authors examine the effect of different choices of the estimator of the covariance matrix for the Wald tests, and different formulations of the density of ε.

6.4.3 Unspecified Alternatives

The Brant (1990) test developed in Section 6.1.2 is ostensibly a test against the null hypothesis,

$$H_0: \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \ldots \boldsymbol{\beta}_{J-1},$$

in the model

$$\text{Prob}(y_i > j \mid \mathbf{x}_i) = F(\boldsymbol{\beta}_j'\mathbf{x}_i - \mu_j), j = 0,\ldots,J-1.$$

The apparently natural alternative hypothesis is the generalized ordered choice model (1). However, that model is not an internally consistent model for the probabilities associated with the outcomes. The presumed alternative does not prevent negative probabilities. One might conclude that the alternative is the "generalized" model when all $\mathbf{x}$'s are such that the probabilities *are* positive – that is, in a certain range of $\mathbf{x}$. However, that range also depends on $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. So, the suggestion amounts to concluding that the model is internally consistent when it is internally consistent. There is no other way to delineate when the model is internally consistent, other than it is when it is. On the other hand, it is persuasive that that Brant test, when it rejects the "null" hypothesis, is picking up some failure of the assumptions of the model. We have examined a variety of generalizations of the ordered choice model; it seems reasonable to conclude that the Brant test might well be finding any of them as an alternative to the base case. Thus, the Brant test might reasonably be considered in the same light as other conditional moment tests. That is, under the null hypothesis, certain features should be observed (within sampling variability). The alternative is, essentially, "not the null."

Butler and Chatterjee (1995, 1997) have reconsidered estimation of the ordered probit model using the generalized method of moments. The null model implies a set of orthogonality conditions based on the definition of the model,

$$E[m_{ij} - (F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i))] = 0,$$

where $m_{ij} = 1$ if $y_i = j$ and 0 otherwise. This provides a set of orthogonality conditions,

$$E\{\mathbf{x}_i[m_{ij} - (F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)]\}, j = 0,\ldots,J.$$

In principle, this implies $(J+1)K$ moment conditions, but one, the last, is redundant. The implied sample moments are, then,

$$\bar{\mathbf{g}}(\boldsymbol{\beta},\boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta},\boldsymbol{\mu})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i[m_{i0} - F(-\boldsymbol{\beta}'\mathbf{x}_i)] \\ \mathbf{x}_i[m_{i1} - [F(\mu_1 - \boldsymbol{\beta}'\mathbf{x}_i) - F(-\boldsymbol{\beta}'\mathbf{x}_i)]] \\ \vdots \\ \mathbf{x}_i[m_{i,J-1} - [F(\mu_{J-1} - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{J-2} - \boldsymbol{\beta}'\mathbf{x}_i)]] \end{pmatrix}.$$

The GMM estimator is then obtained by two steps: (1) Obtain a consistent estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, say the MLE. then compute an estimator of Asy.Var[$\bar{\mathbf{g}}(\boldsymbol{\beta},\boldsymbol{\mu})$], such as

$$\mathbf{V} = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta},\boldsymbol{\mu})\mathbf{g}_i(\boldsymbol{\beta},\boldsymbol{\mu})'\right]$$

(2) minimize the GMM criterion

$$nq = N\bar{\mathbf{g}}(\boldsymbol{\beta},\boldsymbol{\mu})' \mathbf{V}^{-1} \bar{\mathbf{g}}(\boldsymbol{\beta},\boldsymbol{\mu}).$$

The minimized value has a limiting chi squared distribution with degrees of freedom equal to the number of overidentifying restrictions. In this case, the number of moment conditions is $J \times K$ and the number of parameters is $K+J-1$. The number of overidentifying restrictions is $(J-1)(K-1)$. The authors go on to explore the corresponding computations for a bivariate ordered probit model. This proliferates moment conditions, as there is a K-order condition for each pairing of y_{i1} and y_{i2} – though the paucity of observations in some cells might suggest dropping some of the moments. In all cases, it is uncertain what the alternative hypothesis should be if nq is significant. [It is not in the application studied in their paper – see Butler and Chatterjee (1995).] Two suggestions are exogeneity of the independent variabilities and, of course, the distributional assumption.

7

Accommodating Individual Heterogeneity

The presence or absence of individual heterogeneity not contained explicitly in the model is likely to be the most fundamental difference between the bioassay and social science applications of ordered choice models. In the analysis of a population of fruit flies or aphids, the analyst is probably safe in assuming that the population is homogeneous enough to treat with a zero mean, homoscedastic disturbance in the latent tolerance equation and single parameter, and homogeneous thresholds in the observation mechanism. The analysis of a population of congressional representatives or heads of households responding to a survey about health satisfaction or subjective well being will be far from that situation. Consider, as well, the fundamental difference in the underlying equation. For a simple insecticide experiment, the implied underlying regression will be

$$Tolerance_{ir}^{*} = \alpha + \beta\, Treatment_{ir} + \varepsilon_{ir},$$

where i indicates a group (treatment level) and r indicates a member of that group. The entire "behavioral" aspect of the model is embedded in the random term, the "tolerance" to the treatment. The ordered "choice" is

$$\begin{aligned} y_{ir} = \quad & 0 \text{ if } (Tolerance_{ir} - \alpha - \beta\, Treatment_{ir}) \leq \alpha_1 && \text{(dead)}, \\ & 1 \text{ if } (\alpha_1 < Tolerance_{ir} - \alpha - \beta\, Treatment_{ir} \leq \alpha_2) && \text{(moribund)}, \\ & 2 \text{ if } (\alpha_2 < Tolerance_{ir} - \alpha - \beta\, Treatment_{ir}) && \text{(alive)}. \end{aligned}$$

It seems safe to assume that the individual observations are sufficiently homogeneous in dimensions that one could hope to measure that the simple, canonical model above is an adequate description of the outcome variable that we will ultimately observe. In contrast, for the *subjective* well being (SWB) application, the right hand side of the behavioral equation will include variables such as *Income*, *Education*, *Marital Status*, *Children*, *Working Status*, *Health*, and a host of other measurable and unmeasurable, and *measured* and *unmeasured* variables. In individual level behavioral models, such as

$$SWB_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},$$

the relevant question is whether a zero mean, homoscedastic ε_{it}, can be expected to satisfactorily accommodate the likely amount of heterogeneity in the underlying data, and whether it is reasonable to assume that the same thresholds should aply to each individual.

Beginning with Terza (1985), analysts have questioned the adequacy of the ordered choice model from this perspective. As shown below, many of the proposed extensions of the model, such as heteroscedasticity, parameter heterogeneity, etc., parallel developments in other modeling contexts (such as binary choice modeling and modeling counts such as number of doctor visits or hospital visits). The regression based ordered choice model analyzed here does have a unique feature, that the thresholds are part of the behavioral specification. This aspect of the specifiction has been considered as well.

7.1 Threshold Models – The Generalized Ordered Probit Model (2)

The model analyzed thus far assumes that the thresholds μ_j are the same for every individual in the sample. Terza (1985), Pudney and Shields (2000), Boes and Winkelmann (2006a), Greene, Harris, Hollingsworth and Maitra (2008) and Greene and Hensher (2009), all present cases that suggest individual variation in the set of thresholds is a degree of heterogeneity that is likely to be present in the data, but is not accommodated in the model. A precursor to this literature is Farewell (1982), who proposes an ordered Weibull model,

$$\text{Prob}(y_i > j \mid \mathbf{x}_i) = \exp(-\exp(\theta_{ij} - \boldsymbol{\beta}'\mathbf{x}_i)).$$

To accommodate the possibility of latent heterogeneity, he suggests

$$\theta_{ij} = \theta_j^* + \eta_i,$$

with $\theta_{i0} = 0$, so that the spacing between thresholds is preserved, but the location of the set of thresholds varies across individuals. The extreme value functional form is unique. However, the shift of the thresholds points toward the later generalizations of the model, beginning with Terza (1985).

Terza's (1985) generalization of the model is equivalent to

$$\mu_{ij} = \mu_j + \boldsymbol{\delta}'\mathbf{z}_i. \tag{7.1}$$

This is the special case of the generalized model used in his application – his fully general case allows $\boldsymbol{\delta}$ to differ across outcomes. The model is reformulated later to assume that the $\mathbf{z}_i$ in the equation for the thresholds is the same as the $\mathbf{x}_i$ in the regression. For the moment, it is convenient to remove the constant term from $\mathbf{x}_i$. In Terza's application, in which there were three outcomes,

$$y_i^* = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,$$

and

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* \leq 0, \\ &\quad 1 \text{ if } 0 < y_i^* \leq \mu + \boldsymbol{\delta}'\mathbf{x}_i, \\ &\quad 2 \text{ if } y_i^* > \mu + \boldsymbol{\delta}'\mathbf{x}_i. \end{aligned} \tag{7.2}$$

There is an ambiguity in the model as specified. In principle, the model for three outcomes has two thresholds, μ_0 and μ_1. It is always necessary to normalize the first, $\mu_0 = 0$. Therefore, the model implies the following probabilities:

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= \Phi(-\alpha - \boldsymbol{\beta}'\mathbf{x}) &&= 1 - \Phi(\alpha_0 + \boldsymbol{\beta}_0'\mathbf{x}), \\ \text{Prob}(y = 1|\mathbf{x}) &= \Phi(\mu + \boldsymbol{\delta}'\mathbf{x}_i - \alpha - \boldsymbol{\beta}'\mathbf{x}) - \Phi(-\alpha - \boldsymbol{\beta}'\mathbf{x}) &&= \Phi(\alpha_0 + \boldsymbol{\beta}_0'\mathbf{x}) - \Phi(\alpha_1 + \boldsymbol{\beta}_1'\mathbf{x}), \\ \text{Prob}(y = 2|\mathbf{x}) &= \Phi(\alpha + \boldsymbol{\beta}'\mathbf{x} - \mu - \boldsymbol{\delta}'\mathbf{x}) &&= \Phi(\alpha_1 + \boldsymbol{\beta}_1'\mathbf{x}), \end{aligned} \tag{7.3}$$

where $\alpha_0 = \alpha$, $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$, $\alpha_1 = \alpha - \mu$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta} - \boldsymbol{\delta})$. This is precisely Williams's (2006) "Generalized Ordered Probit Model." That is, at this juncture, Terza's heterogeneous thresholds model and the generalized ordered probit model are indistinguishable. For direct applications of Terza's approach, see, e.g., Kerkhofs and Lindeboom (1995), Groot and van den Brink (1999) and Lindeboom and van Doorslayer (2003).

The result carries over generically to the generalized ordered logit and probit models examined earlier. The motivation in these earlier instances, was to work around the parallel regressions assumption. The model specified is

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = F(\mu_j - \boldsymbol{\beta}_j'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}_{j-1}'\mathbf{x}_i). \quad (7.4))$$

Ostensibly, the generalization is to allow a different parameter vector for each outcome. Boes and Winkelmann (2006a, 2006b) proposed the same model, motivated by the single crossing feature of the restricted model. But, when the regressor vector is the same in each cell, the implied "generalized threshold model"

$$\mu_{ij} = \mu_j + \boldsymbol{\gamma}_j'\mathbf{x}_i. \quad (7.5)$$

is also indistinguishable from the model with an outcome specific parameter vector;

$$\boldsymbol{\beta}_j = \boldsymbol{\gamma}_j - \boldsymbol{\beta}. \quad (7.6)$$

We can deduce a comparison of the two models from Terza's results. Terza reports results for a model with five regressors, $\mathbf{x}$ = (CFIE,LTIA,NIIA,TA,CVIA). The numerical results in Table 7.1 are reported in the article (reported estimated standard errors are omitted):

Table 7.1 Estimated Generalized Ordered Probit Models from Terza (1985)

	Ordered Probit	Generalized Ordered Probit		Sample Mean
	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$	$\boldsymbol{\delta}$	
Constant	-2.779	-17.862	-28.617	1.000
x_1	0.604	1.305	2.831	3.069
x_2	3.642	17.788	11.007	0.447
x_3	16.079	124.518	167.130	0.056
x_4	0.0012	0.0007	0.0009	1490.762
x_5	2.865	3.893	10.282	0.176
[μ]	[1.955]	[2.419]		

The estimated value of μ is not reported, but we should be able to approximate it. The sample consists of 222 observations in which the sample counts are 39, 100, 83, so the proportions are P_0 = 0.176, P_1 = 0.450, P_2 = 0.374, respectively. For the middle cell, at least approximately, at the means of the data, we should have,

$$P_1 \approx \Phi[\mu - (a + \mathbf{b}'\bar{\mathbf{x}})] - \Phi[-(a + \mathbf{b}'\bar{\mathbf{x}})].$$

The index function evaluated at the means is approximately 2.026. Using 0.45 for P_1 and the inverse normal function, we obtain a value of μ of approximately 1.955. The log likelihood values are not reported, so it is not possible to compare the two models directly. In the generalized model, the index function evaluated at the means is 2.796. Note that the coefficients have changed wildly; the second has increased by a factor of 4 and the third by a factor of 10. However, when we compute these at the sample means of the data, we find the index function is 2.796 compared to 2.026 previously, and the implied threshold value is 2.419 compared to 1.955. Thus, the changes in the model are fairly moderate. The three predicted probabilities evaluated at the means are (.021382,.450315,.528302) for the first model, and (.002587,.340501,.646909) for the second. (The model would not impose that these mimic the sample, even at the means, as it would in a multinomial (unordered) logit model, so these differences from the sample proportions are to be expected.) The very large swings in the parameter estimates attest to the need to use partial effects to scale them for comparisons across models.

Terza notes (on p. 6) that the model formulation does not impose an ordering on the threshold coefficients. He suggests an inequality constrained maximization of the log likelihood, which is likely to be extremely difficult if there are many variables in **x**. As a "less rigorous but apparently effective remedy," he proposes to drop from the model variables in the threshold equations that are insignificant in the initial (unconstrained) model.

The analysis of this model continues with Pudney and Shields's (2000) "Generalized Ordered Probit Model," [also " Generalized Model (2)"] whose motivation, like Terza's was to accommodate *observable* individual heterogeneity in the threshold parameters as well as in the mean of the regression. (Pudney and Shields studied an example in the context of job promotion in which the steps on the promotion ladder for nurses are somewhat individual specific. In their setting, in contrast to Terza's, at least some of the variables in the threshold equations are explicitly different from those in the regression. Their model (using primarily their notation and their equation numbering in brackets) is a latent regression for career potential

$$y_i^* = \mathbf{x}_i\boldsymbol{\alpha} + \varepsilon_i, \qquad [1]$$

where $\mathbf{x}_i$ is a vector of personal attributes and $\varepsilon_i|\mathbf{x}_i \sim N[0,1]$. Waiting time until promotion, t_i, is affected by career interruptions and potential, y_i^*. Thus, for the observed rank,

$$y_i = g \text{ if } C_{g-1}(y_i^*,\mathbf{q}_i) < t_i \leq C_g(y_i^*,\mathbf{q}_i) \qquad [2]$$

where $\mathbf{q}_i$ is a set of variables that shift the promotion thresholds. $C_g(y_i^*,\mathbf{q}_i)$ are the required waiting times with $C_0 = 0$ and C_m (the top level) = $+\infty$. Since (it is argued) waiting time is monotonic in y_i^*,

$$y_i = g \text{ if } C_{g-1}(t_i,\mathbf{q}_i) < y_i^* \leq C_g(t_i,\mathbf{q}_i) \qquad [3]$$

"The quantities are a nondecreasing sequence of thresholds from which the grade probabilities can be constructed as follows:

$$\Pr(y_i = g|\mathbf{x}_i,\mathbf{q}_i,t_i) = \Phi(C_g - \mathbf{x}_i\boldsymbol{\alpha}) - \Phi(C_{g-1} - \mathbf{x}_i\boldsymbol{\alpha}). \qquad [4]$$

(p. 371). Note that "nondecreasing" above must actually be "increasing" in order to prevent attaching zero probabilities to nonnull events. They then conclude, "observe from equation (4) that, if a variable influences both the promotion thresholds and latent potential y_i^*, then these two influences cannot, in general, be separated." But, this is not true in general; it is true when both y_i^* and C_g are linear functions of t_i and $\mathbf{q}_i$. We pursue this in the next section. The effect of waiting time, t_i is entered into the model by adding linear and quadratic terms in $\psi(t_i) = \min(1,1/t_i)$.

The authors then construct the generalized model and a test of "threshold constancy" by redefining $\mathbf{q}_i$ to be a constant term and those variables that are unique to the threshold model in (2). Variables that are common to both the original $\mathbf{q}_i$ and $\mathbf{x}_i$ are moved in the specification to the regression equation, and the model is reparameterized as

$$\Pr(y_i = g|\mathbf{x}_i,\mathbf{q}_i,t_i) = \Phi[\mathbf{q}_i\boldsymbol{\beta}_g - \mathbf{x}_i(\boldsymbol{\alpha} + \boldsymbol{\delta}_g)] - \Phi[\mathbf{q}_i\boldsymbol{\beta}_{g-1} - \mathbf{x}_i(\boldsymbol{\alpha} + \boldsymbol{\delta}_{g-1})]. \qquad (7.7)$$

The resulting equation is now a hybrid of Terza's and Williams's generalized models, with outcome varying parameters in both thresholds and in the regression. [See (7.6) earlier.] The test of threshold constancy is then carried out simply by testing (using an LM test) the null hypothesis that $\boldsymbol{\delta}_g = \mathbf{0}$ for all $\mathbf{g}$. (A normalization, $\boldsymbol{\delta}_0 = \boldsymbol{\delta}_m = \mathbf{0}$, is imposed at the outset.)

Pudney and Shields' treatment underscores the mathematical equivalence of the varying thresholds model and the "nonparallel regressions" model. Two features of their model to be noted are: First, the probabilities in their revised log likelihood [their equation (8)], are not constrained to be positive. This is obvious from (7.7) Second, the thresholds, $\mathbf{q}_i\boldsymbol{\beta}_g$, are not constrained to be ordered. No restriction on $\boldsymbol{\beta}_g$ will ensure that $\mathbf{q}_i\boldsymbol{\beta}_g > \mathbf{q}_i\boldsymbol{\beta}_{g-1}$ for all data vectors $\mathbf{q}_i$.

The equivalence of the Terza and Williams models is only a mathematical means to the end of estimation of the model. The Pudney and Shields model, itself, has constant parameters in the regression model and outcome varying parameters in the thresholds. They do note, however, (using a more generic notation) a deeper problem of identification). However it is originally formulated, the model implies that

$$\text{Prob}[y_i \leq j \mid \mathbf{x}_i,\mathbf{z}_i] = F(\mu_j + \boldsymbol{\delta}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_i) = F[\mu_j - (\boldsymbol{\delta}^{*\prime}\mathbf{z}_i + \boldsymbol{\beta}'\mathbf{x}_i)], \boldsymbol{\delta}^* = -\boldsymbol{\delta}. \tag{7.8}$$

In their specification, they had a well defined distinction between the variables, $\mathbf{z}_i$ that should appear only in the thresholds and $\mathbf{x}_i$ that should appear in the regression. More generally, it is less than obvious whether the variables $\mathbf{z}_i$ are actually in the threshold or in the mean of the regression. Either interpretation is consistent with the estimable model. Pudney and Shields argue that the distinction is of no substantive consequence for their analysis. The consequence is at the theoretical end, not in the implementation. But, this entire development is necessitated by the linear specification of the thresholds. Absent that, most of the preceding construction is of limited relevance. A more general nonlinear model is discussed in the next section.

7.2 Nonlinear Specifications – A Hierarchical Ordered Probit (HOPIT) Model

The linearity of the regression specification has presented two significant obstacles to building the model. It has rendered indistinguishable the heterogeneous thresholds case and the "generalized" model that has heterogeneous parameter vectors. Second, it has produced a model that will be internally inconsistent at least for some data vectors; that is, it cannot ensure that the probabilities are always positive. One might consider modifying the thresholds directly. Greene (2007a), Eluru, Bhat and Hensher (2008) and Greene and Hensher (2009) propose a "Hierarchical Ordered Probit" or HOPIT Model,

$$\begin{aligned}
&y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\\
&y_i = j \text{ if } \mu_{i,j-1} \leq y_i^* < \mu_{ij},\\
&\mu_0 = 0,\\
&\mu_{i,j} = \exp(\lambda_j + \boldsymbol{\gamma}'\mathbf{z}_i) \quad \text{[Case 1]},\\
\text{or}\quad &\mu_{i,j} = \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i) \quad \text{[Case 2]}.
\end{aligned} \tag{7.9}$$

[The choice of the term "Hierarchical" model might be unfortunate, as it conflicts with a large literature on random parameter models such as the one discussed in Section 8.1, which is a "Hierarchical" model in the sense used in that literature. See, e.g., Raudenbush and Bryk (2002).] Note that case 2 is the Terza(1985) and Pudney and Shields (2000) model with the exponential rather than linear function for the thresholds. It is, however, strongly distinct from Williams's model. This formulation addresses two problems; (i) the thresholds are mathematically distinct from the regression; (ii) by this construction, the threshold parameters must be positive. With a slight modification, to be pursued later, the ordering of the thresholds can also be assured. For the first case,

$$\mu_{i,j} = [\exp(\lambda_1) + \exp(\lambda_2) + ... + \exp(\lambda_j)] \times \exp(\boldsymbol{\gamma}'\mathbf{z}_i), \tag{7.10}$$

and, for the second,

$$\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i). \tag{7.11}$$

In practical terms, the model can now be fit with the constraint that all predicted probabilities are greater than zero. This is a numerical solution to the problem of ordering the thresholds for all data vectors.

This extension of the ordered choice model shows a case of *identification through functional form*. The model parameters $[\lambda_j,\boldsymbol{\gamma}_j,\boldsymbol{\beta}]$ would not be separately identified if all functions were linear. The contemporary literature views, with some skepticism, models that are unidentified without a change in functional form, such as shown above. [See, e.g., King et al. (2004, p. 299).] On the other hand, while this is true, it is also true that the underlying theory of the model does not insist on linearity of the thresholds (or the regression model, for that matter), but it does insist on the ordering of the thresholds, and one might equally criticize the original model for being unidentified *because the model builder insists on a linear form*. That is, there is no obvious reason that the threshold parameters must be linear functions of the variables, or that linearity enjoys some claim to first precedence in the regression function. Of course, this is a methodological issue that cannot be resolved here. [In a similar connection, much of the discussion in Cameron and Heckman (1998), Carniero, Hansen and Heckman (2003), Heckman and Navarro (2005, 2007) and Cunha, Heckman and Navarro (2007) focuses on questions of identification in ordered choice models with random thresholds of the form $c_s(Q_s,\eta_s) = c_s(Q_s) + \eta_s$ where η_s is the random term (their notation). The reason given at the outset that the thresholds must be additive in the random term is to "preserve the separability of the classical ordered choice model." There are cases cited by the authors, such as the role of the tax brackets in a model of labor supply, that might mandate separability. But, as a general rule, this seems like an unnecessary straightjacket. The nonlinearity of the preceding specification, or others that might resemble it, provides the benefit of a simple way to achieve other fundamental results listed by the same authors, e.g., coherency of the model (all positive probabilities).

The partial effects in this model are more involved than have been considered thus far. The Case 2 model implies

$$\text{Prob}(y = j|\mathbf{x},\mathbf{z}) = F(\mu_j - \boldsymbol{\beta}'\mathbf{x}) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}).$$

Thus,

$$\begin{aligned}\frac{\partial \text{Prob}(y = j \mid \mathbf{x},\mathbf{z})}{\partial \mathbf{x}} &= \left[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}) \right]\boldsymbol{\beta}, \\ \frac{\partial \text{Prob}(y = j \mid \mathbf{x},\mathbf{z})}{\partial \mathbf{z}} &= f(\mu_j - \boldsymbol{\beta}'\mathbf{x})\mu_j\boldsymbol{\gamma}_j - f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x})\mu_{j-1}\boldsymbol{\gamma}_{j-1}.\end{aligned} \tag{7.12}$$

(An obvious restriction is imposed if Case 1 applies.) If a variable appears in both $\mathbf{x}$ and $\mathbf{z}$, then the two effects are added. It is clear on inspection, that this formulation has also circumvented the parallel regressions restriction, the single crossing feature and, with separate $\boldsymbol{\delta}$ vectors, the restriction that ratios of partial effects be the same for all outcomes. We conclude that at least as regards the question of functional form, the assumption of linearity has imposed a heavy cost on the construction of the model.

Numerically, the formulation shares the problem that its predecessors have. Without constraints or the modification suggested earlier, it does not impose the ordering of the threshold parameters. Thus, in the general form, unordered thresholds remain a possibility. We have found that the problem seems not to arise very often. As before, starting the iterations at the basic

ordered probit or logit model estimates begins the process with a model in which all probabilities are positive. (At the starting values, $\lambda_j = \log \mu_j$ from the simple model.) As the iterations move the parameters away from the starting values, estimates that move the probabilities toward the proscribed regions begin to impose a heavy penalty on the log likelihood. As before, this appears generally to characterize the optimization process – it is, of course, not a prescription for how to carry it out. We do note, it places a large value on a search method with a sensitive line search – a crude method such as Newton's method (which uses none) is likely to fail early on.

Table 7.2 presents estimates of the ordered probit models using the same formulation as we used earlier. We have modeled the thresholds in terms of *INCOME*, *AGE* and *HANDDUM*, the latter being a dummy variable that indicates whether the individual reports a physical handicap. The table at the top of the listing shows that each succesive generalization of the model brings a significant improvement in the log likelihood – the hypothesis of the restrictions of the preceding model is decisively rejected in all three cases (even if the significance level is adjusted for the sequential testing procedure). This seems consistent with the results found earlier for the Generalized (1) model. There is also a sizable increase (50%) in the Pseudo-R^2, which we will explore in Table 7.4. The estimated coefficients in the index function seem to be relatively stable, save for the coefficient on *INCOME*, which increases substantially as the restrictions of the model are relaxed. This is consistent with the findings reported by Boes and Winkelmann (2006a). It is a bit less surprising when we recall that our data are drawn from the same data base, the GSOEP, as theirs. We may well be examining some of the same individuals.

Table 7.3 displays the partial effects for the three estimated models. Partial effects for the two binary variables that are marked with "*" are computed by discrete changes in the probabilities with other variables held at their means. The effects are strikingly stable in spite of the changes in the coefficients from one model to the next. Table 7.4 suggests the payoff to the generalization. The prediction is the most probable cell computed at the individual observation. The counts of correct predictions for each model are shown in boldface/underline in the table. The effect of the generalization as one moves from left to right is to predict fewer values with $y = 2$ correctly, but more with $y = 1$, and the difference is more than compensated. This would not predict the increase in the pseudo R^2 seen in Table 7.2, but it is consistent with it.

Table 7.2 Estimated Hierarchical Ordered Probit Models

	No Model	Ordered Probit	HO-Case 1	HO-Case 2	
Log likelihood function	-5875.096	-5752.985	-5690.804	-5665.088	
Degrees of Freedom		5	3	6	
Chi squared test of restr.	0	244.222	124.362	51.342	
Info. Criterion: AIC	2.62284	2.57059	2.54419	2.53540	
McFadden Pseudo R-squared	0.00000	.0207847	.0313684	.0357455	

	HOPIT Case 1 Model				HOPIT Case 2 Model				Ordered
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Probit
Constant	2.0300	.1591	12.760	.0000	1.9365	.1869	10.363	.0000	1.9788
AGE	-.0218	.0025	-8.570	.0000	-.0212	.0031	-6.795	.0000	-.0181
EDUC	.0344	.0074	4.656	.0000	.0340	.0074	4.599	.0000	.0356
INCOME	.7349	.1501	4.897	.0000	.9432	.1734	5.440	.0000	.2587
MARRIED	-.0435	.0408	-1.066	.2863	-.0458	.0410	-1.117	.2640	-.0310
KIDS	.0545	.0389	1.402	.1608	.0509	.0390	1.306	.1915	.0606
	μ(j)=exp[λ(j)+δ*z]				μ(j)=exp[λ(j)+γ(j)*z]				
AGE	-.0039	.0011	-3.602	.0003	-.0057	.0025	-2.287	.0222	
INCOME	.2830	.0572	4.952	.0000	.5487	.1222	4.492	.0000	
HANDDUM	.3248	.0235	13.817	.0000	.5058	.0421	12.007	.0000	
AGE					-.0022	.0013	-1.649	.0992	
INCOME					.3152	.0685	4.599	.0000	
HANDDUM					.2578	.0350	7.366	.0000	
AGE					-.0045	.0012	-3.817	.0001	
INCOME					.3353	.0573	5.851	.0000	
HANDDUM					.1805	.0407	4.440	.0000	
	λ(j) in μ(j)=exp(λ(j)+γ′z)				λ(j) in μ(j)=exp(λ(j)+γ(j)′*z				
λ (1)	.1950	.0619	3.151	.0016	.1512	.1296	1.167	.2432	1.1484
λ (2)	.9905	.0551	17.989	.0000	.9100	.0678	13.431	.0000	2.5478
λ (3)	1.1723	.0542	21.634	.0000	1.1837	.0589	20.098	.0000	3.0564

Table 7.3. Estimated Partial Effects for Ordered Probit Models

Variable	Y=00	Y=01	Y=02	Y=03	Y=04
Ordered Probit Model					
AGE	.0017	.0045	-.0012	-.0022	-.0028
EDUC	-.0034	-.0089	.0024	.0042	.0056
INCOME	-.0248	-.0644	.0177	.0309	.0406
*MARRIED	.0029	.0077	-.0020	-.0037	-.0049
*KIDS	-.0057	-.0151	.0040	.0072	.0096
Hierarchical Ordered Probit Model: Case 1					
AGE	.0020	.0055	-.0016	-.0026	-.0032
EDUC	-.0031	-.0087	.0026	.0042	.0051
INCOME	-.0669	-.1860	.0548	.0888	.1093
*MARRIED	.0039	.0110	-.0030	-.0053	-.0066
*KIDS	-.0049	-.0138	.0039	.0066	.0082
Hierarchical Ordered Probit Model: Case 1					
AGE	.0019	.0053	-.0015	-.0024	-.0034
EDUC	-.0031	-.0085	.0024	.0038	.0054
INCOME	-.0861	-.2363	.0666	.1065	.1493
*MARRIED	.0041	.0115	-.0030	-.0052	-.0074
*KIDS	-.0046	-.0127	.0035	.0058	.0081

Table 7.4 Predicted Outcomes from Ordered Probit Models

Cross tabulation of predictions. Row is actual, column is predicted. Predicted Outcome is the one with the largest probability.														
Model		0	1	1	0	1	2	0	1	2	0	1	2	0	1	2
Actual	Row Sum	y=0			y=1			y=2			y=3			y=4		
0	230	**0**	**0**	**0**	60	107		230	170	123	0	0	0	0	0	0
1	1113	0	0	0	**0**	**112**	**215**	1113	1001	898	0	0	0	0	0	0
2	2226	0	0	0	0	84	149	**2226**	**2142**	**2077**	0	0	0	0	0	0
3	500	0	0	0	0	2	10	500	498	490	**0**	**0**	**0**	0	0	0
4	414	0	0	0	0	5	10	414	409	404	0	0	0	**0**	**0**	**0**
Col Sum	4483	0	0	0	263	491		4483	4220	3992	0	0	0	0	0	0

7.3 Thresholds and Heterogeneity – Anchoring Vignettes

The introduction of observed heterogeneity into the threshold parameters attempts to deal with a fundamental assumption of the ordered choice model. Save for the effect of observable heterogeneity just considered, survey respondents view the survey questions essentially the same way. King, Murray, Salomon and Tandon (KMST, 2004) identify two very basic features of survey data that will make this problematic; first, they often measure concepts that are definable only with reference to examples, such as freedom, health, satisfaction, etc. Second, individuals do, in fact, often understand survey questions very differently, particularly with respect to answers at the extremes. A widely used term for this interpersonal incomparability is *differential item functioning* (DIF). Kapteyn, Smith and Van Soest (KSV, 2007) [and Van Soest, Delaney, Harmon, Kapteyn and Smith (VDHKS, 2007)] suggest the results in Figure 7.1 to describe the implications of DIF. The figure shows the distribution of Health (or drinking behavior in the latter study) in two hypothetical countries. The density for country A is to the left of that for country B implying that on average, those in country A are less healthy than those in country B. But, the people in the two countries use very different response scales if asked to report their health on a five point scale as shown. In the figure, those in country A have a much more positive view of a given health status than those in country B. A person in country A with health status indicated by the dotted line would report that they are in "Very Good" health while a person in country B with the same health status would report only "Fair." A simple frequency of the distribution of self-assessments of health status in the two countries would suggest that people in country A are much healthier than those in country B when, in fact, the opposite is true. Correcting for the influences of DIF in such a situation would be essential to obtaining a meaningful comparison of the two countries. The impact of DIF is an accepted feature of the model within a population, but could be strongly distortionary when comparing very disparate groups, such as across countries, as in KMST (2004, political groups), Murray, Tandon, Mathers and Sudana (2002, health outcomes), Tandon et al. (2004), KSV (2007, work disability), Sirven, Santos-Egglmann and Spagnoli (2008) and Gupta, Kristensena and Possoli (2008, health), all of whom used the ordered probit model to make cross group comparisons. Other recent applications include Angelina et al. (2008, life satisfaction), Kristensen and Johansson (2008) and Bago d'Uva et al. (2008).

KMST proposed the use of *anchoring vignettes* to resolve this difference in perceptions across groups. The essential approach is to use a series of examples, the same for all respondents, to estimate each respondent's DIF and correct for it. [The idea of using vignettes to anchor perceptions in survey questions is not itself new; KMST cite a number of earlier uses. The innovation here is their method for incorporating the approach in a formal model for the ordered choices.] Consider their example. The self assessment is of *political efficacy*; "How much say do you have in getting the government to address issues that interest you?" A set of

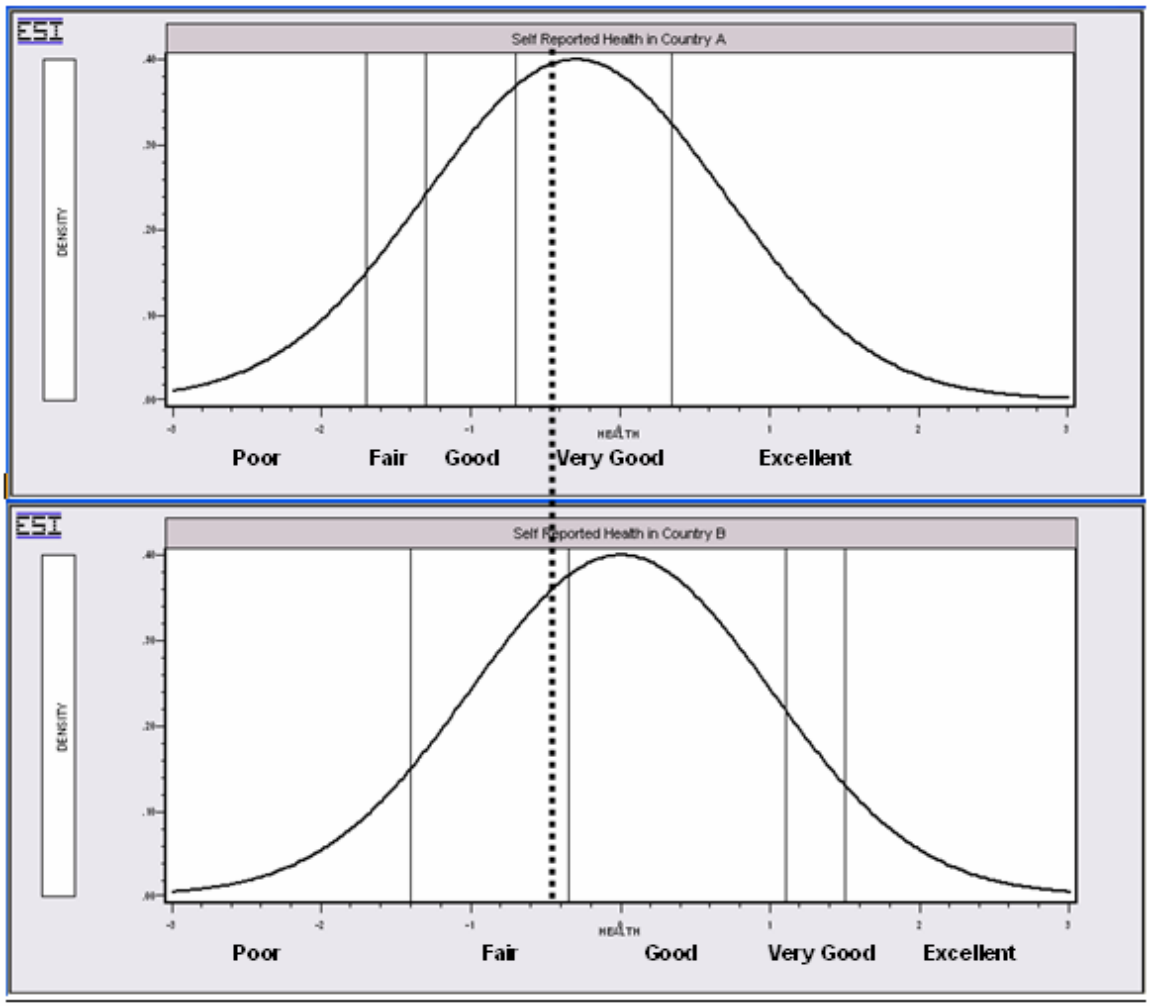


Figure 7.1 Differential Item Functioning in Ordered Choices

ordinal response categories is offered: (1) None, (2) Little, (3) Some, (4) A lot, (5) Unlimited. A set of vignettes, is posed with the same responses:

1. *[Allison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.*
2. *[Imelda] lacks clean drinking water. She and her neighbors are drawing attention to the issue by collecting signatures on a petition. They plan to present the petition to each of the political parties before the upcoming election.*
3. *[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it will be futile to vote for the opposition party since the government is certain to win.*
4. *[Toshiro] lacks clean drinking water. There is a group of local leaders who could do something about the problem, but they have said that inductrial development is the most important policy right now instead of clean water.*
5. *[Moses] lacks clean drinking water. He would like to change this but he can't vote and feels that no one in the government cares about this issue. So he suffers in silence, hoping that something will be done in the future.*

The vignettes fall on an ordered scale from most to least efficacy. The same question as the self-assessment is asked with respect to the person in each vignette. The vignette questions address a specific dimension of political efficacy. The analysis assumes that the self assessment addresses the same (whatever) concept of political efficacy as the vignettes. KMST suggest that the self-assessment be asked first, followed by the vignettes randomly ordered and suitably renamed to match culture and gender where possible.

Three issues now seem pertinent:

(1) Where does the analyst obtain the vignettes? There is a thriving literature on vignettes. Three rich sources are King and Wand (2007), Hopkins and King (2008) and King (2008). For another example, KSV (2007, p. 465, fn 6) provide a URL for the vignettes used in their study. King (2009) lists many additional sources in a large number of fields within the social sciences.

(2) What assumptions are needed to make this a viable approach to handling DIF? The authors list two key assumptions:

> *Vignette equivalence* is the assumption that the level of the variable represented by a particular vignette is perceived by all respondents in the same way and on the same scale apart from random measurement error. (This is what makes them *anchoring* vignettes.) Differences across individuals are assumed to be random with respect to the characteristic being measured. (The cross country difference shown in the example in Figure 7.1 violates this assumption.)
>
> *Response consistency* is the assumption that each individual uses the response categories for a particular survey question in the same way when providing a self-assessment as when assessing each of the hypothetical people in the vignettes. There can be heterogeneity (DIF) across individuals, and across different questions (with their vignettes) within a survey. But, it is assumed that

there is no DIF within the the group of items defined by a single question with its vignettes for a particular person

The model formulation developed by KMST and others embodies parametric restrictions that correspond to these two assumptions.

(3) How does the analyst use the information? KMST propose a nonparametric approach that shows visually how the use of anchoring vignettes can calibrate the responses to a common base. They then develop a formal extension of the ordered probit model, which we will describe in detail. [See, also, Tandon et al. (2004).] Software for estimation of the model include routines in Stata and NLOGIT, and packages described by King and Wand (2007) and Wand, King and Lau (2007).

7.3.1 Using Anchoring Vignettes in the Ordered Probit Model

The specification that incorporates the anchoring vignettes consists of two ordered choice models, the *self assessment component* and the *vignette component*. The sample data and notation used for them are as follows:

Self-assessment:
There are N individuals surveyed denoted $i = 1,...,N$.
The self assessment consists of S questions, indexed $s = 1,...,S$.
Each question is answered with J possibilities, $j = 1,...,J$. These may vary across questions, however, for simplicity, we will assume not.

Vignettes:
There are Q individuals posed the vignette questions. These need not be the same individuals posed the self-assessment questions, and Q need not equal N.
Individuals posed the vignettes are indexed $q = 1,...,Q$.
There are M vignette questions, $m = 1,...,M$.
Each vignette is answered with J_1 possibilities, $j = 1,...,J_1$, the same responses as for the first self-assessment.

Self Assessment Component

We begin by assuming that $S = 1$; a single self-assessment question is posed. The case of multiple self assessment questions is treated below. The self assessment component begins with the usual latent regression;

$$y_i^* = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \varepsilon_i \sim N[0,1]. \quad (7.13)$$

The heterogeneity and the ordering of the thresholds are imposed as follows:

$$\begin{aligned}
&\mu_{i,0} = -\infty, \mu_{i,J} = +\infty \\
&\mu_{i,1} = \lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z}_i \ , \\
&\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i), j = 2,...,J\text{-}1, \quad (7.14)
\end{aligned}$$

(KMST assume that there is no overall constant term in the latent regression. We have added one here to maintain consistency with the model in Section 7.2. We have also isolated the constant terms in the thresholds. We will reconcile the two formulations shortly.) The question is answered on a J point scale, $j = 1,...,J$. The measurement

equation is

$$y_i = j \text{ if } \mu_{i,j-1} < y_i^* \leq \mu_{i,j}, j = 1,\ldots,J. \tag{7.15}$$

The assumptions about $\mu_{i,0}$, $\mu_{i,J}$ and σ_ε^2 are the usual location and scaling restrictions.

As shown in the Appendix to this chapter, this model is mathematically identical to the HOPIT model in Section 7.2, Case 2 in (7.9) earlier;

$$\begin{aligned} y_i^* &= \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - (\lambda_0 + \boldsymbol{\gamma}_0'\mathbf{z}_i) + \varepsilon_i \\ \tau_{i,-1} &= -\infty,\ \tau_{i,0} = 0,\ \tau_{i,J} = +\infty. \\ \tau_{i,j} &= \tau_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i), j = 1,\ldots J\text{-}1 \\ y_i &= j \text{ if } \tau_{i,j-1} < y_i^* \leq \tau_{i,j}, j = 0,\ldots,J. \end{aligned} \tag{7.16}$$

(The accommodation of our earlier observation scheme, $y_i = 0,1,\ldots,J$, rather than from 1 to J, now requires only a trivial change in notation.) The self assessment model thus accommodates the individual heterogeneity in the thresholds by introducing $\mathbf{z}_i$ in $\tau_{i,j}$.

The reconstruction of (7.13 and 7.14) as (7.16) highlights a now familiar problem of identification – because the first threshold in the KMST model is specified linearly, if $\mathbf{x}_i$ and $\mathbf{z}_i$ have variables in common, then the respective parts of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_0$ cannot be separately estimated. This is true of the constant term as well. KMST (p. 299) note: "Response category DIF appears in the model as threshold variation (τ_{ij} and τ_{lj} varying over respondents i and l) and requires at least one vignette for strong identification. We can see the essential role of vignettes by what happens if we try to estimate the self-assessment component separately and, also, set the explanatory variables X affecting the actual level to be the same as those V affecting the thresholds. In this case, $\boldsymbol{\beta}$ (the effect of X) and $\boldsymbol{\gamma}$ (the effect of V) would be dubiously identified only from the nonlinearities in the threshold model (5)." Once again, we note that nonlinearity is deemed dubious, while linearity is not. However, in fact, the parts of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_0$ that correspond to variables that are in both $\mathbf{x}_i$ and $\mathbf{z}_i$ are not separately identified at all in the absence of the vignettes. This shows up in the constant term as well. The result is unrelated to the nonlinearities in $\tau_{i,j}$. However, it is true that only the nonlinearity of $\tau_{i,j}$ identifies $(\lambda_j,\boldsymbol{\gamma}_j)$ for $j > 1$ when $\mathbf{x}_i = \mathbf{z}_i$.

The model parameters to be estimated are the same as in the model in Section 7.2, $(\beta_0,\boldsymbol{\beta},\lambda_0,\boldsymbol{\gamma}_0),\lambda_1,\boldsymbol{\gamma}_1,\ldots,\lambda_{J-1},\boldsymbol{\gamma}_{J-1})$ subject to the indeterminacies such as the constant term, $\alpha_0 = (\beta_0 - \lambda_0)$. The contribution of the self assessment component to the log likelihood is precisely that of case (2) of the hierarchical ordered probit model in Section 7.2;

$$\log L_A = \sum_{i=1}^{N} \log \sum_{j=0}^{J} B_{i,j} \left[\begin{array}{l} \Phi\left(\tau_{i,j} - (\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_0 - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) - \\ \Phi\left(\tau_{i,j-1} - (\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_0 - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) \end{array} \right], \tag{7.17}$$

where $\tau_{i,j}$ is defined in (7.16) and B_{ij} equals 1 if $y_i = j$ and 0 otherwise. Again, it is clear that as it stands, the likelihood function does not produce separate estimates of β_0 and λ_0 nor of any components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_0$ for which the corresponding variables are the same. These would be estimated as $\alpha_0 = (\beta_0 - \lambda_0)$ and $\alpha_k = (\beta_k - \gamma_{0,k})$, respectively. Looking ahead, this lack of identifiability is resolved by a cross equation restriction between the assessment and the vignette models. The remaining parameters, $(\lambda_j,\boldsymbol{\gamma}_j,j=1,\ldots,J\text{-}1)$ are, as noted by KMST, identified through the nonlinearity of $\tau_{i,j}$.

Vignette Component

The vignette component of the model is formulated as follows: Denote the actual level for the hypothetical person described in vignette m as θ_m, $m = 1,...,M$. The assumption of vignette equivalence is imposed by assuming that θ_m is the same for all individuals. Then, for the sample of people asked the vignettes the latent regression is

$$w_{m,q}^* = \theta_m + \sigma e_{mq} \text{ where } e_{mq} \sim N[0,1]. \tag{7.18}$$

The vignette is assumed to have the same set of responses as the first self-assessment question. Responses to the vignette question are formed by

$$\begin{aligned} &w_{m,q}^* = \theta_m - (\lambda_0 + \boldsymbol{\gamma}_0'\mathbf{z}_m) + \sigma e_{mq}, \\ &\tau_{-1,q} = -\infty,\ \tau_{0,q} = 0,\ \tau_{J,q} = +\infty, \\ &\tau_{j,q} = \tau_{j-1,q} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_q), j = 1,...J\text{-}1 \\ &w_{m,q} = j \text{ if } \tau_{j-1,q} < w_{m,q}^* < \tau_{j,q}, j = 0,\ldots,J. \end{aligned} \tag{7.19}$$

The thresholds are restricted to match those for the first self-assessment. The anchoring aspect of the vignette questions is obtained by the constraint that the same $(\lambda_j,\boldsymbol{\gamma}_j)$ appears in the vignette model and the self-assessment model. (We have gone directly to the reparameterization consistent with the HOPIT model of Section 7.2.)

The contribution of the vignettes sample to the log likelihood is

$$\log L_V = \sum_{q=1}^{Q} \sum_{m=1}^{M} \log \sum_{j=0}^{J} B_{q,j,m} \left[\begin{array}{l} \Phi\left\{\frac{1}{\sigma}\left[\tau_{j,q} - (\theta_m - (\lambda_0 + \boldsymbol{\gamma}_0'\mathbf{z}_q))\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma}\left[\tau_{j-1,q} - (\theta_m - (\lambda_0 + \boldsymbol{\gamma}_0'\mathbf{z}_q))\right]\right\} \end{array} \right]. \tag{7.20}$$

The new parameters to be estimated are $(\theta_1,...,\theta_M,\sigma)$. No new parameters are introduced by the thresholds in the vignette component; they are the same as appear in the self-assessment model. In cases considered thus far, it was not possible to estimate a scale parameter in the ordered choice model. The parameter σ is identified in the vignette model because there are numerous other restrictions. The threshold parameters are the same in all M equations. With two or more vignettes, it is possible to estimate an unrestricted σ. This is essentially the same form of identification that allows estimation of a variance parameter in the random effects model in Section 9.2.

KSV (2007) suggest a refinement of the model by introducing covariates in the latent regression in (7.17). In their application, they are interested in work disability, and propose to add gender (FEMALE) to the model. The vignette model, with all components becomes

$$\begin{aligned} &w_{m,q}^* = \theta_m + \boldsymbol{\delta}_m'\mathbf{a}_q - (\lambda_0 + \boldsymbol{\gamma}_0'\mathbf{z}_m) + \sigma e_{mq} \\ &\tau_{-1,m} = -\infty,\ \tau_{0,m} = 0,\ \tau_{J,m} = +\infty, \\ &\tau_{j,m} = \tau_{j-1,m,q} + \exp(\lambda_{j,m} + \boldsymbol{\gamma}_{j,m}'\mathbf{z}_q), j = 1,...J\text{-}1 \\ &w_{m,q} = j \text{ if } \tau_{j-1,m} < w_{m,q}^* < \tau_{j,m}, j = 0,\ldots,J. \end{aligned}$$

where $\boldsymbol{\delta}_m$ is the new parameter vector and $\mathbf{a}_q$ is the vector of covariates.

In their study of drinking behavior, VDHKS have used the same model as KSV with a sex dummy variable and a dummy variable for whether the respondent was shown a definition of a drink before the interview. They treated the multiple vignettes as "repeated measures," and added a common random effect, u_q, to each of the vignette regressions.

7.3.2 Log Likelihood and Model Identification Through the Anchoring Vignettes

The log likelihood for the vignette equations could be maximized separately from the log likelihood function for the self-asssessment. The same identification issue as before would seem to arises with respect to the constant terms, $\alpha_m = (\theta_m - \lambda_0)$. However, there are cross equation constraints that the same $(\lambda_j, \boldsymbol{\gamma}_j, j=0,\ldots,J\text{-}1)$ appears in both the self-assessment and vignette log likelihoods. This is the *anchoring vignette restriction.* Thus, the substantive restriction of the model is embodied in this equality restriction for the parameter vector $(\lambda_j, \boldsymbol{\gamma}_j, j=0,\ldots,J\text{-}1)$. We can now see the point of KMST's observation about identification. The parameter vector is (apart from the constant terms, which appear in the form $\eta_m = (\theta_m\text{-}\lambda_0)$), identified by the vignette equations. This means that with at least one vignette, $\boldsymbol{\gamma}_0$ is estimable, which implies that $\boldsymbol{\beta}$ is also. (Once again, this argument falls apart if the threshold parameters are all specified as linear functions.) The crucial result is that each vignette, by itself, produces identification of the first threshold parameters, $(\boldsymbol{\gamma}_0)$. With more than one vignette, the parameters are, in fact, overidentified. However, as long as β_0 is nonzero, λ_0 is never estimable. (Or, vice versa.)

To use all the information in the sample, the log likelihood function is the sum of the two parts, with the restriction on the common threshold parameters,

$$\begin{aligned}\log L = & \sum_{i=1}^{N} \log \sum_{j=0}^{J} B_{i,j}\left[\begin{array}{l}\Phi\left(\tau_{i,j} - (\beta_0-\lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) - \\ \Phi\left(\tau_{i,j-1} - (\beta_0-\lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right)\end{array}\right] \\ & + \sum_{q=1}^{Q}\sum_{m=1}^{M} \log \sum_{j=0}^{J} B_{q,j,m}\left[\begin{array}{l}\Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j} - (\theta_m-\lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j-1} - (\theta_m-\lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\}\end{array}\right]\end{aligned} \tag{7.21}$$

The issue of scaling in the vignette equations is a nontrivial loose end in this strand of literature. KMST note (2004, p. 198), for example, "Although we avoid complicating the notation here, we often let σ^2 vary over vignettes, since their estimates are convenient indicators of how well each vignette is understood." As long as there are sufficient restrictions implied by the common threshold parameters, it will be possible to estimate separate scaling parameters by the vignette model. KSV (2007) specify a self-assessment specific variance, σ_s^2 and a common vignette variance, σ^2, but ultimately normalize all at one with "… can be identified (up to the usual normalization of scale and location)." With sufficient difference between $\mathbf{x}_i$ and $\mathbf{z}_i$, and with at least one vignette, the parameters of the model are overidentified. This will allow separate estimation of vignette specific variances. The log likelihood for KMST's model would be

$$\begin{aligned}\log L = & \sum_{i=1}^{N} \log \sum_{j=0}^{J} B_{i,j}\left[\begin{array}{l}\Phi\left(\tau_{i,j} - (\beta_0-\lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) - \\ \Phi\left(\tau_{i,j-1} - (\beta_0-\lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right)\end{array}\right] \\ & + \sum_{q=1}^{Q}\sum_{m=1}^{M} \log \sum_{j=0}^{J} B_{q,j,m}\left[\begin{array}{l}\Phi\left\{\frac{1}{\sigma_m}\left[\tau_{q,j} - (\theta_m-\lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma_m}\left[\tau_{q,j-1} - (\theta_m-\lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\}\end{array}\right]\end{aligned} \tag{7.22}$$

Recent applications, including VDHKS (2007), KSV. (2007) and Gupta, Kristensen and Pozzoli (2008) have not extended the model in this direction. In spite of the mathematical degree of freedom, it would seem that homoscedasticity of the vignette equations would be a desirable feature of the model.

7.3.3 Testing the Assumptions of the Model

As noted, there are two fundamental assumptions underlying the model, vignette equivalence and response consistency. No formal test is suggested for the first of these. KMST (p. 199) suggest that an informal test for vignette equivalence is suggested by the vignette equations. Under the equivalence assumption, in the original model (7.18), the θ_m values should be ordered. However, they offer a number of other explanations for a finding that the estimates are not ranked. This approach is only suggestive. Moreover, this "test" loses it's appeal if there are covariates in the vignette equations, as in KSV (2007).

Response consistency is imposed by assuming that all the threshold parameters, are common to the two models and across vignettes. This is not a directly testable restriction. GKR assert "In order to identify separate thresholds in the subjective self-reports and the vignette evaluations, we need more information – with the subjective self-reports and the vignette evaluations alone, identification requires the maintained assumption of response consistency." Their formal test is constructed by adding a third, "objective" measure of the concept measured by the subjective self-assessment,

$$
\begin{aligned}
&O_i^* = \beta_0^O + \boldsymbol{\beta}^{O\prime}\mathbf{x}_i + \varepsilon_{i,o} \\
&\mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty. \\
&\mu_j = \mu_{j-1} + \exp(\chi_j), j = 1,...J\text{-}1 \\
&O_i = j \text{ if } \mu_{j-1} < O_i^* \leq \mu_j, j = 0,\ldots,J.
\end{aligned}
\tag{7.23}
$$

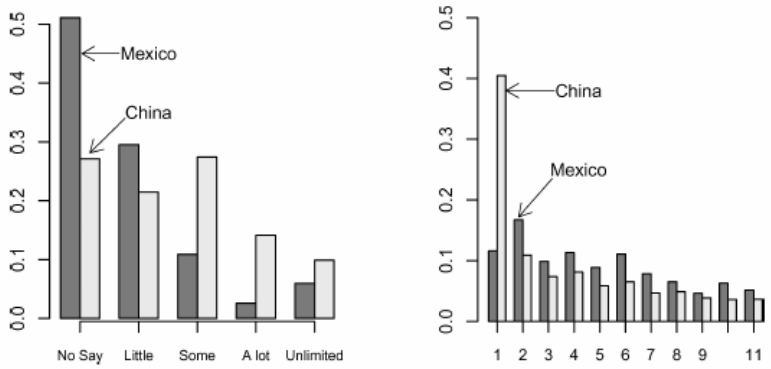
In VDHKS's (2007) application, the subjective measure of student drinking behavior was "*How would you describe your own drinking patterns over the course of the last year?*" *Mild, Moderate, Some Cause for Concern, Excessive/Extreme*. The objective measure quantified the actual number of drinks consumed on a day that the student was drinking. In Gupta, et al.'s study of health measured across several European countries, the objective assessment was a measure of hand grip strength.

The "one factor" assumption is $(\beta_0^O, \boldsymbol{\beta}^O) = (\beta_0, \boldsymbol{\beta})$. The objective assessment adds an equation to the model. Under null hypothesis of the one factor model, the restriction will serve to identify $(\beta_0, \boldsymbol{\beta})$ in the self-assessment model, and the response consistency assumption is no longer needed. One can then relax the response consistency assumption – the vignettes and subjective self assessments will identify their threshold parameters and the mean parameters in the threshold latent regressions. The mean parameters in the subjective self-assessment equation, β_0 and $\boldsymbol{\beta}$ that were previously mixed with λ_0 and $\boldsymbol{\gamma}_0$ are now separately identified by the objective assessment equation, leaving λ_0 and $\boldsymbol{\gamma}_0$ separately identified in the mean of the self-assessment equation. The authors explore several approaches to testing different aspects of the model.

7.3.4 Application

Figure 7.2 is Figure 2 from KMST (2004). The authors describe two assessments of political efficacy based on surveys of individuals in Mexico and China. The left panel shows the raw tallies, and suggests, counter to intuition, that individuals in China have a substantially greater assessment of their political effiicacy than people in Mexico. The right panel shows a nonparametric approach to the modeling described here that strikingly reverses the conclusion. Figure 7.3 shows their estimates of an ordered probit model for self-assessed efficacy with vignettes. The ordered probit results in the center are consistent with the left panel in Figure 7.2. Of particular interest is the coefficient on the "China" dummy variable that suggests that Chinese individuals are much more likely to answer with the highest level than Mexican citizens are. The rightmost two columns report estimates of their "CHOPIT" model that reverses the conclusion. (KMST provide an extensive analysis of the empirical results in their study.)

FIGURE 2. Nonparametric Estimates of an Electoral Dimension of Political Efficacy



Note: The left graph is a histogram of the observed categorical self-assessments. The right graph is a histogram of *C*, our nonparametric DIF-corrected estimate of the same distribution.

Figure 7.2 KMST Comparison of Political Efficacy

TABLE 2. Comparing Political Efficacy in Mexico and China

		Ordered Probit		Our Method	
Eq.	Variable	Coeff.	(SE)	Coeff.	(SE)
μ	**China**	**0.670**	(**0.082**)	**−0.364**	(**0.090**)
	Age	0.004	(0.003)	0.006	(0.003)
	Male	0.087	(0.076)	0.114	(0.081)
	Education	0.020	(0.008)	0.020	(0.008)
τ^1	**China**			**−1.059**	(**0.059**)
	Age			0.002	(0.001)
	Male			0.044	(0.036)
	Education			−0.001	(0.004)
	Constant	0.425	(0.147)	0.431	(0.151)
τ^2	**China**			**−0.162**	(**0.071**)
	Age			−0.002	(0.002)
	Male			−0.059	(0.051)
	Education			0.001	(0.006)
	Constant	−0.320	(0.059)	−0.245	(0.114)
τ^3	**China**			**0.345**	(**0.053**)
	Age			−0.001	(0.002)
	Male			0.044	(0.047)
	Education			−0.003	(0.005)
	Constant	−0.449	(0.074)	−0.476	(0.105)
τ^4	**China**			**0.631**	(**0.083**)
	Age			0.004	(0.002)
	Male			−0.097	(0.072)
	Education			0.027	(0.007)
	Constant	−0.898	(0.119)	−1.621	(0.149)
Vignettes	θ_1			1.284	(0.161)
	θ_2			1.196	(0.160)
	θ_3			0.845	(0.159)
	θ_4			0.795	(0.159)
	θ_5			0.621	(0.159)
$\ln \sigma$				−0.239	(0.042)

Note: Ordered probit indicates counterintuitively and probably incorrectly that the Chinese have higher political efficacy than the Mexicans, whereas our approach reveals that this is because the Chinese have comparatively lower standards (τ's) for moving from one categorical response into the next highest category. The result is that although the Chinese give higher reported levels of political efficacy than the Mexicans, it is the Mexicans who are in fact more politically efficacious.

Figure 7.3 KMST Estimated Vignette Model

7.3.5 Multiple Self-Assessment Equations

If more than one self assessment question is posed, then the self assessment component of KMST's model takes the form of the random effects model developed in Section 9.2. For the *s*th self-assessment,

$$
\begin{aligned}
&y_{i,s}^* = (\beta_0 - \lambda_0) + \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i + \varepsilon_{i,s} + u_i, \\
&\tau_{-1,s} = -\infty,\ \tau_{0,s} = 0,\ \tau_{J,} = +\infty, \\
&\tau_{j,s} = \tau_{j-1,s} + \exp(\lambda_{j,s} + \boldsymbol{\gamma}_{j,s}'\mathbf{z}_i), j = 1,\ldots J\text{-}1, \\
&y_{i,s} = j \text{ if } \tau_{i,j-1,s} < y_{i,s}^* \leq \tau_{i,j,s},\ j = 0,\ldots,J.
\end{aligned}
\tag{7.24}
$$

The log likelihood for this expanded form of the model will be

$$
\begin{aligned}
\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \prod_{s=1}^{S} \left\{ \sum_{j=0}^{J} B_{i,j,s} \left[\begin{array}{l} \Phi\left(\tau_{i,j,s} - ((\beta_0 - \lambda_0) + (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{v}_i)) - \sigma c_i\right) - \\ \Phi\left(\tau_{i,j-1,s} - ((\beta_0 - \lambda_0) + (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{v}_i)) - \sigma c_i\right) \end{array} \right] \right\} \phi(c_i) dc_i \\
+ \sum_{q=1}^{Q} \sum_{m=1}^{M} \log \left\{ \sum_{j=0}^{J} B_{q,j,m} \left[\begin{array}{l} \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j,1} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma}\left[\mu_{q,j-1,1} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} \end{array} \right] \right\}
\end{aligned}
\tag{7.25}
$$

Further details on computation of a random effects ordered probit model appear in Section 9.2.

There is some nontrivial disagreement in the received applications over the treatment of the random effects. In KMST's placement of the heterogeneity, u_i (above), it appears as a random effect only in the self-assessment component of the model. Later treatments, e.g., KSV. (2007), Van Soest et al. (2007), Gupta et al. (2008), have placed the common effect in the first threshold. In the original parameterization, they write

$$
\begin{aligned}
&\mu_{i,0} = -\infty,\ \mu_{i,J} = +\infty \\
&\mu_{i,1} = \lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z}_i + \mathrm{u_i} \\
&\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i), j = 2,\ldots,J\text{-}1,.
\end{aligned}
$$

Since the thresholds are common to all of the equations, the presence of u_i in both parts of the model would imply that the log likelihood is not separable as in (7.25). It also implies, in contrast to KMST that the vignette questions and the self-assessment questions must be posed to the same sample of individuals. The appropriate log likelihood by this construction would be formed as

$$
\begin{aligned}
L_{S,i} \mid c_i &= \prod_{s=1}^{S} \prod_{j=0}^{J} \left[\begin{array}{l} \Phi\left(\tau_{i,j,s} - ((\beta_0 - \lambda_0) + (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{v}_i)) - \sigma c_i\right) - \\ \Phi\left(\tau_{i,j-1,s} - ((\beta_0 - \lambda_0) + (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{v}_i)) - \sigma c_i\right) \end{array} \right]^{B_{i,j,s}} \\
L_{V,i} \mid c_i &= \prod_{m=1}^{M} \prod_{j=0}^{J} \left[\begin{array}{l} \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j,1} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q - \sigma c_i\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j-1,1} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q - \sigma c_i\right]\right\} \end{array} \right]^{B_{i,j,m}} \\
\log L &= \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \left(L_{S,i} \mid c_i\right)\left(L_{V,i} \mid c_i\right) \phi(c_i) dc_i
\end{aligned}
\tag{7.26}
$$

The difference is more than mathematical. KMST include u_i as a random effect in the mean of the latent regression for the self-assessment equation. KSV state, instead, "The term u_i introduces

an unobserved individual effect in the response scale. It implies that evaluations of different vignettes are correlated with each other and with the self-reports (conditional on $\mathbf{x}_i$), since some respondents will tend to use high thresholds and others will use low thresholds in all their evaluations." The observation seems appropriate, however, one might question whether the variance within the set of vignettes or the correlation across them should be identical to that across the self-assessments. A form of heteroscedasticity,

$$\sigma_{u,i}^2 = \sigma^2 \times \exp(1 + \kappa V_i) \tag{7.27}$$

where V_i is a dummy variable for whether the question is a vignette or a self-assessment, would seem a natural extension of the model.

7.4 Heterogeneous Scaling (Heteroscedasticity) of Random Utility

Considerably less attention has been focused on specification of the conditional variance in the regression model than on the conditional mean and the thresholds. In microeconomic data, scaling of the underlying preferences is surely as important a source of heterogeneity as displacement of the mean, perhaps even more so. One would expect the problem of heterogeneity of the variance to be a persistent feature of individual level data. Researchers questioned its implications as early as Cox (1970). [See, also, Cox (1995).] Nonetheless, formal treatment of the issue is a relatively recent extension of the model.

A heteroscedastic ordered choice model is a minor extension of the basic model; the following form of the model based on Harvey (1976) appears in earlier versions of *LIMDEP* [Econometric Software (1997)], *Stata* [Stata, Version 8] and in Bhat (1999) as a natural extension of the binary probit and logit models. The ordered choice model with heteroscedasticity would be

$$\begin{aligned}
y_i^* &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \\
y_i &= 0 \text{ if } \mu_{-1} < y_i^* \le \mu_0, \\
&= 1 \text{ if } \mu_0 < y_i^* \le \mu_1, \\
&= 2 \text{ if } \mu_1 < y_i^* \le \mu_2 \\
&= \ldots \\
&= J \text{ if } \mu_{J-1} < y_i^* \le \mu_J, \\
\text{Var}[\varepsilon_i|\mathbf{h}_i] &\propto [\exp(\boldsymbol{\gamma}'\mathbf{h}_i)]^2.
\end{aligned} \tag{7,28}$$

The model is also discussed in some detail in Williams (2006) and is a feature of `GOLogit` and `GOProbit`. A search of the literature will turn up hundreds of recent applications of binary and ordered choice models with this form of heteroscedasticity [e.g., Hensher (2006)]. The binary probit and logit models with this form of heteroscedasticity are obvious extensions of the basic probit model, and appear much earlier, e.g., in Greene (1990) and Allison (1999).

Recall, at the outset of the discussion, it emerged that the lack of information on scaling of ε and therefore y^* is a signature feature of the ordered choice model. This same result will have major implications for building heteroscedasticity into the model. Consider the formulation of the model used in Chen and Khan (2003),

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + [\exp(\boldsymbol{\gamma}'\mathbf{h}_i)]\varepsilon_i, \tag{7.29}$$

where ε_i is still N[0,1]. It follows that the observation mechanism is now

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{h}_i) = F\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right) - F\left(\frac{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right). \tag{7.30}$$

This straightforward extension of the model should bring a substantive improvement in the correspondence of the model to the underlying data. Greene (2007a) proposes to blend this model with the hierarchical model of the Section 7.2. The resulting functional form,

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{h}_i) = F\left(\frac{\exp(\theta_j + \boldsymbol{\delta}_j'\mathbf{z}_i) - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right) - F\left(\frac{\exp(\theta_{j-1} + \boldsymbol{\delta}_{j-1}'\mathbf{z}_i) - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right), \tag{7.31}$$

should be intricate enough to overcome the parallel regressions, single crossing and constant ratios features of the basic model.

Unlike the linear regression case, unaccounted for heteroscedasticity is potentially disastrous for estimation of the parameters in the ordered choice model. In the presence of latent heteroscedasticity that involves the variables that are in the model, or variables that are correlated with the variables in the model, the maximum likelihood estimator will be inconsistent, potentially seriously so. It is easy to see why in the formulation above. Unlike the linear regression model, in which latent heteroscedasticity will merely taint the standard errors, in the ordered (and binary) choice model, it will masquerade as a change in the functional form. Consider the model above, which can be written in equivalent form

$$y_i^{**} = \boldsymbol{\beta}'\mathbf{x}_i/[\exp(\boldsymbol{\gamma}'\mathbf{h}_i)] + \varepsilon_i,$$

$$y_i = j \text{ if } \mu_{i,j} \leq y_i^{**} < \mu_{i,j}, \tag{7.32}$$

where $\varepsilon_i \sim N[0,1]$,

but $\mu_{i,j} = \mu_j/[\exp(\boldsymbol{\gamma}'\mathbf{h}_i)]$.

That is, the equivalent form of the model is one with a highly nonlinear conditional mean function and heterogeneous thresholds. Recall, the data contain no independent information on scaling of the underlying variable – any such information is determined from the conditional means and the functional form adopted for the variance. Estimating the model as if the disturbance were homoscedastic ignores both of these facts. Note that computing a "robust" covariance matrix for the estimator does nothing to redeem it. The estimator is inconsistent, so the robust covariance matrix estimator is a moot point. Keele and Park (2005) have examined this model and its implications for bias in estimation. Chen and Khan (2003) have reconsidered the estimation of this model using robust methods that allow estimation of $\boldsymbol{\beta}$ even in the presence of heteroscedasticity. But, estimation of $\boldsymbol{\beta}$ solves only part of the model builder's problem. If the measured outcome takes more than three values, then partial effects will be required to make much sense of the estimates. Without information about the underlying variance, or the underlying distribution, the scaling needed for the transformation is not computable.

As in other cases, the modification of the model alters the partial effects. For this case (omitting the hierarchical probit effects), the marginal effects are

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{h}_i)}{\partial \mathbf{x}_i} = \left[f\left(\frac{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right) - f\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right)\right] \exp(-\boldsymbol{\gamma}'\mathbf{h}_i)\boldsymbol{\beta},$$

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{h}_i)}{\partial \mathbf{h}_i} = \left[\begin{array}{l} f\left(\frac{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right)\left(\frac{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right) \\ -f\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right)\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i)}\right) \end{array} \right] \boldsymbol{\gamma}. \quad . \qquad (7.33)$$

For a variable that appears in both $\mathbf{x}_i$ and $\mathbf{h}_i$, the two parts are added. In such a case, the interpretation of the element of $\boldsymbol{\beta}$ associated with a particular variable becomes even more ambiguous than before.

Table 7.5 displays the estimates of the heteroscedastic ordered probit model using our earlier specification but adding *INCOME*, *AGE* and gender (*FEMALE*) to the variance equation. The basic slope parameters are quite similar to the earlier model (shown in the right panel of Table 7.5 for convenience) But, the evidence of heteroscedasticity with respect to age and income is statistically significant, both individually and using the likelihood ratio test for the larger model. (The value of chi squared with 3 degrees of freedom is 2(5752.985-5741.624) = 22.722. The tabled critical value is 7.814, so on this basis and based on the individual tests, the hypothesis of homoscedasticity would be rejected. It seems likely that this is yet another possible explanation for the finding of the Brant test carried out earlier.

Table 7.5 Estimated Heteroscedastic Ordered Probit Model

Ordered Probability Model								
Dependent variable			HEALTH					
Log likelihood function:	Hetero.		Homosk.					
	-5741.624		-5752.985					
Info. Criterion: AIC:	2.56686		2.57059					
	Heteroscedastic Ordered Probit				Ordered Probit			
	LogL = -5741.624				LogL = -5752.985			
	LogLR = -5752.985				LogL0 = -5875.096			
	Chisq = 22.722				Chisq = 244.2238			
	Degrees of Freeedom 3				Degrees of Freedom 5			
	PseudoRqq = .0227183				PseudoRsq= .0217845			
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	2.1935	.1778	12.337	.0000	1.9788	.1162	17.034	.0000	1.0000
AGE	-.0199	.0021	-9.398	.0000	-.0181	.0016	-11.166	.0000	43.4401
EDUC	.0390	.0080	4.869	.0000	.0356	.0071	4.986	.0000	11.4181
INCOME	.2499	.0863	2.895	.0038	.2587	.1039	2.490	.0128	.34874
MARRIED	-.0306	.0444	-.688	.4916	-.0310	.0420	-.737	.4608	.75217
KIDS	.0698	.0417	1.674	.0942	.0606	.0382	1.586	.1127	.37943
Variance Function									
INCOME	-.2359	.0607	-3.883	.0001					.34874
FEMALE	.0168	.0249	.673	.5009					.48404
AGE	.0037	.0011	3.337	.0008					43.4401
Threshold Parameters									
Mu(1)	1.2817	.0811	15.795	.0000	1.1484	.0212	54.274	.0000	
Mu(2)	2.8019	.1592	17.605	.0000	2.5478	.0216	117.856	.0000	
Mu(3)	3.3507	.1874	17.881	.0000	3.0564	.0267	115.500	.0000	

Table 7.6 displays the partial effects from both the restricted model and the heteroscedastic model. The latter are decomposed into the mean effects ($\partial P(.)/\partial \mathbf{x}$), the variance effects, ($\partial P(.)/\partial \mathbf{h}$) and the total equal to the sum of the two. The parts are marked; the total effects are

shown in boldface. The partial effects from the restricted (homoscedastic) model are shown in parentheses for comparison. In contrast to the raw coefficients, the partial effects have shown some fairly substantial changes. The effects of *AGE* and *INCOME* are quite different (and changes sign twice), while the partial effects for *EDUC* , *MARRIED* and *KIDS* are quite similar to their earlier values .

Table 7.6 Partial Effects in Heteroscedastic Ordered Probit Model

Marginal Effects for Ordered Probit						
Variable	HEALTH=0	HEALTH=1	HEALTH=2	HEALTH=3	HEALTH=4	
AGE	.00169	.00463	-.00128	-.00216	-.00288	Mean
AGE	.00618	.00103	-.01647	.00086	.00839	Variance
AGE	**.00787**	**.00566**	**-.01775**	**-.00130**	**.00551**	**Total**
(AGE)	(.0017)	(.0045)	(-.0012)	(-.0022)	(-.0028)	Restricted
EDUC	-.00332	-.00906	.00251	.00423	.00564	Total
(EDUC)	(-.0034)	(-.0089)	(.0024)	(.0042)	(.0056)	restricted
INCOME	-.02122	-.05800	.01607	.02704	.03611	Mean
INCOME	.34732	.05785	-.92501	.04858	.47126	Variance
INCOME	**.32610**	**-.00015**	**-.90894**	**.07562**	**.50737**	**Total**
(INCOME)	(-.0248)	(-.0644)	(.0177)	(.0309)	(.0406)	Restricted
MARRIED	.00260	.00709	-.00197	-.00331	-.00442	Total
(MARRIED)	(.0029)	(.0077)	(-.0020)	(-.0037)	(-.0049)	Restricted
KIDS	-.00593	-.01620	.00449	.00755	.01008	Total
(KIDS)	(-.0057)	(-.0151))	(.0040)	(.0072)	(.0096)	Restricted
Pure Variance Effect						
FEMALE	-.00316	-.00053	.00840	-.00044	-.00428	Total

7.5 Individually Heterogeneous Marginal Utilities

Greene (2002, 2008a) argues that the fixed parameter version of the ordered choice model (and more generally, many microeconometric specifications) do not adequately account for the underlying heterogeneity likely to be present in observed data. Further extensions of the ordered choice model presented there include full random parameters treatments and discrete approximations under the form of latent class, or finite mixture models. These two specific extensions are also listed by Boes and Winkelmann (2006a).

The preceding lists the received "generalizations" of the ordered choice model. (The many other modified ordered choice models, such as bivariate ordered choice models, models with sample selection, and zero inflation models, that appear elsewhere have not been mentioned, as they are proposed to deal with features of the data other than heterogeneity. We will describe some of them in the chapters to follow.) In what follows, we will propose a formulation of the ordered choice model that relaxes the restrictions listed above but treats heterogeneity in a unified, internally consistent fashion. The model contains three points at which individual heterogeneity can substantively appear, in the random utility model (the marginal utilities), in the threshold parameters, and in the scaling (variance) of the random components. As argued above, this form of treatment seems more likely to capture the salient features of the data generating mechanism than the received "generalized ordered logit model."

Appendix: Equivalence of the Vignette and HOPIT Models

KMST's (2004) model formulation for the self-assessment component is

$$
\begin{aligned}
y_i^* &= \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}[0,1].\\
\mu_{i,0} &= -\infty,\ \mu_{i,J} = +\infty\\
\mu_{i,1} &= \lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z}_i \quad ,\\
\mu_{i,j} &= \mu_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i), j = 2,...,J\text{-}1,\\
y_i &= j \text{ if } \mu_{i,j-1} < y_i^* \le \mu_{i,j}, j = 1,\ldots,J.
\end{aligned}
$$

The authors assume that there is no overall constant term in the latent regression. We have added β_1 to maintain consistency with the earlier model. We have also isolated the constant terms in the thresholds, λ_j – they do include constants in $\tau_{i,j}$ implicitly in the definitions. The definition of $\mu_{i,1}$ and the last (the observation) equation imply

$$
\begin{aligned}
y_i &= 1 \text{ if } -\infty < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \le \mu_{i,1},\\
y_i &= 2 \text{ if } \mu_{i,1} < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \le \mu_{i,1} + \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i),\\
y_i &= j \text{ if } \mu_{i,1} + \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_{j-1} + \boldsymbol{\gamma}_{j-1}'\mathbf{z}_i) < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \le\\
&\qquad \mu_{i,1} + \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i),\\
y_i &= J \text{ if } \mu_{i,1} + \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_{J-1} + \boldsymbol{\gamma}_{J-1}'\mathbf{z}_i) < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \le +\infty.
\end{aligned}
$$

Note that $\mu_{i,1} = \lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z}_i$ appears linearly in every threshold. By subtracting $(\lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z}_i)$ from each finite term, we obtain the equivalent model,

$$
\begin{aligned}
y_i &= 1 \text{ if } -\infty < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_1 - \boldsymbol{\gamma}_1'\mathbf{z} + \varepsilon_i \le 0\\
y_i &= 2 \text{ if } 0 < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_1 - \boldsymbol{\gamma}_1'\mathbf{z} + \varepsilon_i \le \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i)\\
y_i &= j \text{ if } \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_{j-1} + \boldsymbol{\gamma}_{j-1}'\mathbf{z}_i) < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_1 - \boldsymbol{\gamma}_1'\mathbf{z} + \varepsilon_i \le\\
&\qquad \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i)\\
y_i &= J \text{ if } \exp(\lambda_2 + \boldsymbol{\gamma}_2'\mathbf{z}_i) \ldots + \exp(\lambda_{J-1} + \boldsymbol{\gamma}_{J-1}'\mathbf{z}_i) < \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i - \lambda_1 + \boldsymbol{\gamma}_1'\mathbf{z} + \varepsilon_i \le +\infty.
\end{aligned}
$$

This implies the underlying structure

$$
\begin{aligned}
y_i^* &= (\beta_0 - \lambda_1) + \boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_1'\mathbf{z}_i + \varepsilon_i, \varepsilon_i \sim \mathrm{N}[0,1],\\
\tau_{i,0} &= -\infty, \tau_{i,1} = 0, \tau_{i,J} = +\infty.,\\
\tau_{i,j} &= \tau_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i), j = 2,...J\text{-}1\\
y_i &= j \text{ if } \tau_{i,j-1} < y_i^* \le \tau_{i,j}, j = 1,\ldots,J.
\end{aligned}
$$

This is identical to the model in Section 7.2 (subject to some identification problems for the constant term and coefficients in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_1$ that multiply the same variables). Accommodating the coding convention $y_i = 0,1,\ldots,J$ rather than 1 to J is done with a trivial change in notation in the model above. The "j" subscriptes on λ_j, $\boldsymbol{\gamma}_j$ and $\tau_{i,j}$ are reduced by one, and $j = 0,1,\ldots,J$.

8

Parameter Variation and a Generalized Ordered Choice Model

Formal modeling of heterogeneity in the parameters as representing a feature of the underlying data, appears in Greene (2002) (version 8.0), Bhat (1999), Bhat and Zhao (2002) and Boes and Winkelmann (2006). These treatments suggest a full random parameters (RP) approach to the model. In Boes and Winkelmann, however, it is noted that the nature of an RP specification induces heteroscedasticity, and could simply be modeled as such.

8.1 Random Parameters Models

In the same fashion as Swamy's (1970,1971) treatment of random parameters in the linear regression model, one approach to accommodating random parameters is to construct the reduced form of the basic model. This produces a model with heteroscedasticity. We briefly examine this approach, then turn to the structural approach that characterizes the contemporary methods.

8.1.1 Implied Heteroscedasticity

Boes and Winkelmann's (2006) treatment of a zero constant term and a full set of threshold parameters will prove less convenient than including a constant in $\mathbf{x}_i$ and setting $\mu_0 = 0$, instead. We will maintain the latter formulation. The model would appear as follows:

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

where $\mathbf{u}_i \sim N[\mathbf{0},\boldsymbol{\Omega}]$. Inserting the expression for $\boldsymbol{\beta}_i$ in the latent regression model, we obtain

$$\begin{aligned} y_i^* &= \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i \\ &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i + \mathbf{x}_i'\mathbf{u}_i. \end{aligned}$$

The observation mechanism is the same as earlier. The result is an ordered probit model in which the disturbance has variance $\text{Var}[\varepsilon_i + \mathbf{x}_i'\mathbf{u}_i] = 1 + \mathbf{x}_i'\boldsymbol{\Omega}\mathbf{x}_i$; that is, a heteroscedastic ordered probit model. The resulting model has

$$\text{Prob}[y_i \le j \mid \mathbf{x}_i] = \text{Prob}[\varepsilon_i + \mathbf{u}_i'\mathbf{x}_i \le \mu_j - \boldsymbol{\beta}'\mathbf{x}_i] = F\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i}{\sqrt{1 + \mathbf{x}_i'\boldsymbol{\Omega}\mathbf{x}_i}}\right), \tag{8.1}$$

which, it is suggested, can be estimated by ordinary means, albeit with a new source of nonlinearity – the elements of $\boldsymbol{\Omega}$ must now be estimated as well. (The authors' suggestion that this could be handled semiparametrically without specifying a distribution for $\mathbf{u}_i$ is incorrect, because the resulting heteroscedastic ordered choice model as written above only preserves the standard normal form assumed if $\mathbf{u}_i$ is normally distributed as well as ε_i.) They did not pursue this approach. This computation will present a series of difficulties owing to the need to force $\boldsymbol{\Omega}$ to be a positive definite matrix. One cannot simply insert the function above into the log likelihood and be optimistic that the estimated unconstrained matrix will, indeed, stay positive

definite. At worst, it will become indefinite and it will become impossible to compute the log likelihood. A standard remedy is to use a Cholesky decomposition of $\boldsymbol{\Omega}$. Write $\boldsymbol{\Omega} = \mathbf{L}\mathbf{D}^2\mathbf{L}'$ where $\mathbf{D}$ is a diagonal matrix with strictly positive elements and $\mathbf{L}$ is a lower triangular matrix with ones on the diagonal. The log likelihood is then maximized with respect to the elements of $\mathbf{L}$ and $\mathbf{D}$ in addition to $\boldsymbol{\beta}$ and $\mu_1,...,\mu_{J-1}$. This will preserve the positive definiteness of the implied covariance matrix. Elements of $\boldsymbol{\Omega}$ can be deduced after estimation.

Partial effects in this model can be obtained by differentiating the probabilities as if the parts in the numerators and denominators are functions of different variables, then adding them. An expression for this result is given in Boes and Winkelmann (2006a). An application in their study was done under the assumption that $\boldsymbol{\Omega}$ is diagonal ($\mathbf{L} = \mathbf{I}$ in our formulation), which then requires only that the variances of the random parameters be positive.

8.1.2 Maximum Simulated Likelihood Estimation

Greene (2002, 2007a, 2008a,b) analyzes the same model, but estimates the parameters by maximum simulated likelihood. First, write the random parameters as

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \Delta\mathbf{z}_i + \mathbf{L}\mathbf{D}\mathbf{w}_i, \tag{8.2}$$

where $\mathbf{w}_i$ has a multivariate standard normal distribution, and $\mathbf{L}\mathbf{D}^2\mathbf{L}' = \boldsymbol{\Omega}$. The Cholesky matrix, $\mathbf{L}$, is lower triangular with ones on the diagonal. The below diagonal elements of $\mathbf{L}$, λ_{mn}, produce the nonzero correlations across parameters. The diagonal matrix, $\mathbf{D}$, provides the scale factors, δ_m, i.e., the standard deviations of the random parameters. The end result is that $\mathbf{L}(\mathbf{D}\mathbf{w}_i)$ is a mixture, $\mathbf{L}\mathbf{w}_i^*$ of random variables, w_{im}^* which have variances δ_m^2. This is a two level 'hierarchical' model (in the more widely used sense). The probability for an observation is

$$\begin{aligned}\text{Prob}(y_i=j|\mathbf{x}_i,\mathbf{w}_i) &= \left[\Phi\left(\mu_j-\boldsymbol{\beta}_i'\mathbf{x}_i\right)-\Phi\left(\mu_{j-1}-\boldsymbol{\beta}_i'\mathbf{x}_i\right)\right] \\ &= \begin{bmatrix}\Phi\left(\mu_j-\boldsymbol{\beta}'\mathbf{x}_i-\mathbf{z}_i'\Delta'\mathbf{x}_i-(\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right)- \\ \Phi\left(\mu_{j-1}-\boldsymbol{\beta}'\mathbf{x}_i-\mathbf{z}_i'\Delta'\mathbf{x}_i-(\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right)\end{bmatrix}.\end{aligned} \tag{8.3}$$

In order to maximize the log likelihood, we must first integrate out the elements of the unobserved $\mathbf{w}_i$. Thus, the contribution to the unconditional log likelihood for observation i is

$$\log L_i = \log\int_{\mathbf{w}_i}\begin{bmatrix}\Phi\left(\mu_j-\boldsymbol{\beta}'\mathbf{x}_i-\mathbf{z}_i'\Delta'\mathbf{x}_i-(\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right)- \\ \Phi\left(\mu_{j-1}-\boldsymbol{\beta}'\mathbf{x}_i-\mathbf{z}_i'\Delta'\mathbf{x}_i-(\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right)\end{bmatrix}F(\mathbf{w}_i)d\mathbf{w}_i. \tag{8.4}$$

The log likelihood for the sample is then the sum over the observations. Computing the integrals is an obstacle that must now be overcome. It has been simplified considerably already by decomposing $\boldsymbol{\Omega}$ explicitly in the log likelihood, so that $F(\mathbf{w}_i)$ is the multivariate standard normal density. The *Stata* routine, `GLAMM` [Rabe-Hesketh, Skrondal and Pickles (2005)] that is used for some discrete choice models does the computation using a form of Hermite quadrature. An alternative, generally substantially faster method of maximizing the log likelhood is maximum simulated likelihood. The integration is replaced with a simulation over R draws from the multivariate standard normal population. The simulated log likelihood is

$$\log L_S = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \begin{bmatrix} \Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - \mathbf{z}_i'\boldsymbol{\Delta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_{ir})'\mathbf{x}_i\right) - \\ \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - \mathbf{z}_i'\boldsymbol{\Delta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_{ir})'\mathbf{x}_i\right) \end{bmatrix}. \tag{8.5}$$

The simulations are speeded up considerably by using Halton draws [see Halton (1970) for the general principle, and Bhat (2001, 2003) and Train (2003) for applications in the estimation of 'mixed logit models'] rather than random draws. Further details on this method of estimation are also given in Greene (2007b, 2008a). Partial effects and predicted probabilities must be simulated as well. For the partial effects,

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \int_{\mathbf{w}_i} \begin{bmatrix} \phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - \mathbf{z}_i'\boldsymbol{\Delta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right) - \\ \phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - \mathbf{z}_i'\boldsymbol{\Delta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right) \end{bmatrix} (\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i - \mathbf{L}\mathbf{D}\mathbf{w}_i)\, F(\mathbf{w}_i) d\mathbf{w}_i. \tag{8.6}$$

we use simulation to compute

$$Est.\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \left\{ \frac{1}{R} \sum_{r=1}^{R} \begin{bmatrix} \phi\left(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \mathbf{z}_i'\hat{\boldsymbol{\Delta}}'\mathbf{x}_i - (\hat{\mathbf{L}}\hat{\mathbf{D}}\mathbf{w}_{ir})'\mathbf{x}_i\right) - \\ \phi\left(\hat{\mu}_j - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \mathbf{z}_i'\hat{\boldsymbol{\Delta}}'\mathbf{x}_i - (\hat{\mathbf{L}}\hat{\mathbf{D}}\mathbf{w}_{ir})'\mathbf{x}_i\right) \end{bmatrix} \right\} (\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Delta}}\mathbf{z}_i - \hat{\mathbf{L}}\hat{\mathbf{D}}\mathbf{w}_{ir}). \tag{8.7}$$

Table 8.1 gives the estimates of the random parameters model for our familiar specification. The estimator produces estimates of **L** and **D**. The implied estimate of $\boldsymbol{\Omega}$ is given in Table 8.2 with the estimates of the square roots of the diagonal elements of $\boldsymbol{\Omega}$ and the implied correlation matrix obtained by $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}$ where $\hat{\boldsymbol{\Sigma}}$ is a diagonal matrix containing the estimated standard deviations from $\hat{\boldsymbol{\Omega}}$. The estimates of the partial effects are shown in Table 8.3 with their counterparts from the basic model. A likelihood ratio test of the null hypothesis that the basic model applies against the alternative of this generalization is based on a chi squared statistic of 2(5752.985 – 5705.592) = 94.786 with 20 degrees of freedom. The null hypothesis of the model with nonrandom parameters would be rejected.

8.1.3 Variance Heterogeneity

An extension of the model that allows for heterogeneity in the variances of $\boldsymbol{\beta}_i$ as well as the means (in the form of $\boldsymbol{\Delta}\mathbf{z}_i$) is obtained by parameterizing the elements δ_m in **D**. The model with heteroscedasticity is obtained by

$$\mathbf{D}_{im} = \text{diag}[\delta_m \times \exp(\boldsymbol{\eta}_m'\mathbf{a}_i)], \tag{8.8}$$

where $\mathbf{a}_i$ is a vector of covariates specified to act on the variances of the random parameters, rather than the means. This model [implemented in Greene (2002, 2007)] has the potential to proliferate parameters, particularly if there is a nonzero $\boldsymbol{\Delta}$ in the model. One would typically restrict many of the structural parameters to equal zero. In the application in Bhat and Zhao (2002), they use **L** = **I** (no correlation across parameters), $\boldsymbol{\Delta}$ = **0** (no heterogeneity in the means of

Table 8.1 Estimated Random Parameters Ordered Probit Model

Random Coefficients Ordered Probit Model				
Number of observations	4483			
Log likelihood function	-5705.592			
Simulation based on 25 Halton draws				
Variable	**Coefficient**	**Standard Error**	**b/St.Er.**	**P[|Z|>z]**
Means for random parameters				
Constant	3.2142	.1366	23.527	.0000
AGE	-.0298	.0018	-16.379	.0000
EDUC	.0599	.0077	7.760	.0000
INCOME	.5584	.1163	4.800	.0000
MARRIED	-.1140	.0461	-2.476	.0133
KIDS	.1167	.0419	2.785	.0053
Threshold parameters for probabilities				
MU(1)	1.9283	.0519	37.157	.0000
MU(2)	4.1843	.0676	61.928	.0000
MU(3)	5.0018	.0742	67.351	.0000

Table 8.2 Implied Estimates of Parameter Matrices*

	Constant	AGE	EDUC	INCOME	MARRIED	KIDS
Cholesky Matrix L with D on Diagonal						
Constant	1.5902	0	0	0	0	0
AGE	.0081	.0002	0	0	0	0
EDUC	-.0264	-.0203	.0031	0	0	0
INCOME	-1.5621	-1.3753	.9964	.5257	0	0
MARRIED	-.2130	1.2528	.4322	.5738	.1829	0
KIDS	-.7280	.3384	-1.0681	-.6322	-.1446	.2507
$\mathbf{LD^2L'} = \mathbf{W}$ = Implied covariance matrix of random parameters						
Constant	2.5287					
AGE	0.0129	6.55609e-005				
EDUC	-0.0420	-0.0002	0.0011			
INCOME	-2.4841	-0.0129	0.0724	5.6009		
MARRIED	-0.3387	-0.0015	-0.0185	-0.6580	2.1642	
KIDS	-1.1576	-0.0058	0.0090	-0.7249	-0.2718	2.2689
Square roots of diagonal elements						
	1.5902	0.0081	0.0335	2.3666	1.4711	1.5063
Implied correlation matrix of random parameters						
Constant	1.0000					
AGE	0.9997	1.0000				
EDUC	-0.7888	-0.8039	1.0000			
INCOME	-0.6601	-0.6746	0.9128	1.0000		
MARRIED	-0.1448	-0.1232	-0.3760	-0.1890	1.0000	
KIDS	-0.4833	-0.4774	0.1789	-0.2033	-0.1227	1.0000

* Estimated standard errors omitted.

Table 8.3 Estimated Partial Effects from Random Parameters Model

Summary of Marginal Effects for Ordered Probability Models										
Effects computed at means. Effects for binary variables are computed as differences of probabilities, other variables at means.										
	Ordered Probit					Random Parameters Ordered Probit				
Outcome	**AGE**	**EDUC**	**INCOME**	**MARRIED**	**KIDS**	**AGE**	**EDUC**	**INCOME**	**MARRIED**	**KIDS**
Y = 00	.0003	-.0005	-.0050	.0017	-.0010	.0017	-.0034	-.0248	.0029	-.0057
Y = 01	.0081	-.0164	-.1528	.0305	-.0316	.0045	-.0089	-.0644	.0077	-.0151
Y = 02	-.0041	.0083	.0770	-.0143	.0154	-.0012	.0024	.0177	-.0020	.0040
Y = 03	-.0033	.0067	.0627	-.0132	.0133	-.0022	.0042	.0308	-.0037	.0072
Y = 04	-.0010	.0019	.0180	-.0039	.0039	-.0028	.0056	.0406	-.0049	.0096

the random parameters) and specify only one heteroscedastic random parameter, the constant term. This greatly reduces the complexity of the model.

8.1.4 Conditional Mean Estimation in the Random Parameters Model

The random parameters model is couched in terms of $(\boldsymbol{\beta}_i,\mu_1,...,\mu_{J-1})$, specific to the individual. Recall in the structure,

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i.$$

It would be useful to estimate $\boldsymbol{\beta}_i$ rather than the population parameters, $\boldsymbol{\beta}$, if that were possible. It is not, as that would require estimation of $\mathbf{u}_i$ which is "noise." However, in the same spirit as its Bayesian counterpart, one can compute an estimate of $E[\boldsymbol{\beta}_i|y_i,\mathbf{x}_i]$, which will contain more information than the natural, unconditional estimator, $\boldsymbol{\beta}$. The approach proceeds as follows: The density of $y_i|\mathbf{x}_i,\boldsymbol{\beta}_i$ is

$$P(y_i|\mathbf{x}_i,\boldsymbol{\beta}_i) = \text{Prob}(y_i=j|\mathbf{x}_i,\boldsymbol{\beta}_i) = \left[\Phi\left(\mu_j - \boldsymbol{\beta}_i'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}_i'\mathbf{x}_i\right)\right]. \tag{8.9}$$

The marginal density of $\boldsymbol{\beta}_i$ assuming $\mathbf{u}_i \sim \text{N}[\mathbf{0},\boldsymbol{\Omega}]$ is $N[\boldsymbol{\beta},\boldsymbol{\Omega}]$. The joint density of y_i and $\boldsymbol{\beta}_i$ is

$$P(y_i,\boldsymbol{\beta}_i|\mathbf{x}_i) = \text{Prob}(y_i = j\,|\mathbf{x}_i,\boldsymbol{\beta}_i)\,P(\boldsymbol{\beta}_i).$$

Using Bayes Theorem, then,

$$\begin{aligned} P(\boldsymbol{\beta}_i \mid y_i,\mathbf{x}_i) &= \frac{P(y \mid \mathbf{x}_i,\boldsymbol{\beta}_i)P(\boldsymbol{\beta}_i)}{P(y_i)} \\ &= \frac{P(y \mid \mathbf{x}_i,\boldsymbol{\beta}_i)P(\boldsymbol{\beta}_i)}{\int_{\boldsymbol{\beta}_i} P(y_i \mid \mathbf{x}_i,\boldsymbol{\beta}_i)P(\boldsymbol{\beta}_i)d\boldsymbol{\beta}_i}. \end{aligned} \tag{8.10}$$

The conditional mean is then,

$$E(\boldsymbol{\beta}_i \mid y_i,\mathbf{x}_i) = \frac{\int_{\boldsymbol{\beta}_i} \boldsymbol{\beta}_i P(y \mid \mathbf{x}_i,\boldsymbol{\beta}_i)P(\boldsymbol{\beta}_i)d\boldsymbol{\beta}_i}{\int_{\boldsymbol{\beta}_i} P(y_i \mid \mathbf{x}_i,\boldsymbol{\beta}_i)P(\boldsymbol{\beta}_i)d\boldsymbol{\beta}_i}\ . \tag{8.11}$$

The integrals must be computed by simulation. The result is easily obtained as a byproduct of the estimation process. To see how, first insert the components of the probabilities, and replace the integration with simulation, as we did in computing the log likelihood. Then,

$$Est.E(\boldsymbol{\beta}_i \mid y_i,\mathbf{x}_i) = \frac{\frac{1}{R}\sum_{r=1}^{R} \boldsymbol{\beta}_{ir} \sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_j - \boldsymbol{\beta}_{ir}'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}_{ir}'\mathbf{x}_i\right)\right]}{\frac{1}{R}\sum_{r=1}^{R} \sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_j - \boldsymbol{\beta}_{ir}'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}_{ir}'\mathbf{x}_i\right)\right]}\ , \tag{8.12}$$

where $m_{ij} = 1(y_i = j)$. The terms in square brackets are the simulated probabilities that enter the log likelihood. The draws on $\boldsymbol{\beta}_{ir}$ are obtained during the simulation; they are

$$\boldsymbol{\beta}_{ir} = \boldsymbol{\beta} + \mathbf{LDw}_{ir}.$$

That is, the same simulation that was done to maximize the log likelihood. It is illuminating to write this in a different form. Write this, using our final estimates of the model parameters,

$$0 < \hat{a}_{ir} = \frac{\left[\Phi\left(\hat{\mu}_j - \hat{\boldsymbol{\beta}}_{ir}'\mathbf{x}_i\right) - \Phi\left(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}_{ir}'\mathbf{x}_i\right)\right]}{\frac{1}{R}\sum_{r=1}^{R}\left[\Phi\left(\hat{\mu}_j - \hat{\boldsymbol{\beta}}_{ir}'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \hat{\boldsymbol{\beta}}_{ir}'\mathbf{x}_i\right)\right]} < 1\,, \tag{8.13}$$

where

$$\hat{\boldsymbol{\beta}}_{ir} = \hat{\boldsymbol{\beta}} + \hat{\mathbf{L}}\hat{\mathbf{D}}\mathbf{w}_{ir}\,.$$

Then, our estimator is

$$Est.E(\boldsymbol{\beta}_i \mid y_i, \mathbf{x}_i) = \frac{1}{R}\sum_{r=1}^{R} \hat{a}_{ir}\,\hat{\boldsymbol{\beta}}_{ir}\,. \tag{8.14}$$

Other functions of the parameters, such as partial effects or probabilities for individual observations, could be simulated in the same way, just by replacing $\boldsymbol{\beta}_i$ with the desired function of $\boldsymbol{\beta}_i$ in the simulation.

This is not a direct estimator of $\boldsymbol{\beta}_i$; it is an estimator of the mean of the conditional distribution from which $\boldsymbol{\beta}_i$ is drawn. In the classical framework we are using here, this is as well as we can do, in terms of using the sample information, to estimate $\boldsymbol{\beta}_i$. This estimator is a counterpart to the Bayesian posterior mean, which would estimate the same parameters in the same way. A difference would be that the Bayesian posterior variance would be smaller than the variance of the conditional distribution if we computed it above. The reason is that our classical estimator uses the asymptotic distribution of the estimator while the Bayesian posterior mean is conditioned only on (is posterior to) the observed sample. There is a degree of imprecision in the classical estimator that is absent from the posterior mean, because the simulations plug in the estimates of the parameters as if they were known, while the Bayesian counterpart is based on the exact, finite sample distribution of the estimators conditioned on the data in hand. This latter difference is likely to be extremely small in a sample as large as the one in use here. Figure 8.1 shows a kernel density estimator for the distribution of estimates of $E[\beta_{INCOME}|y_{i\cdot},\mathbf{x}_i]$ across the sample.

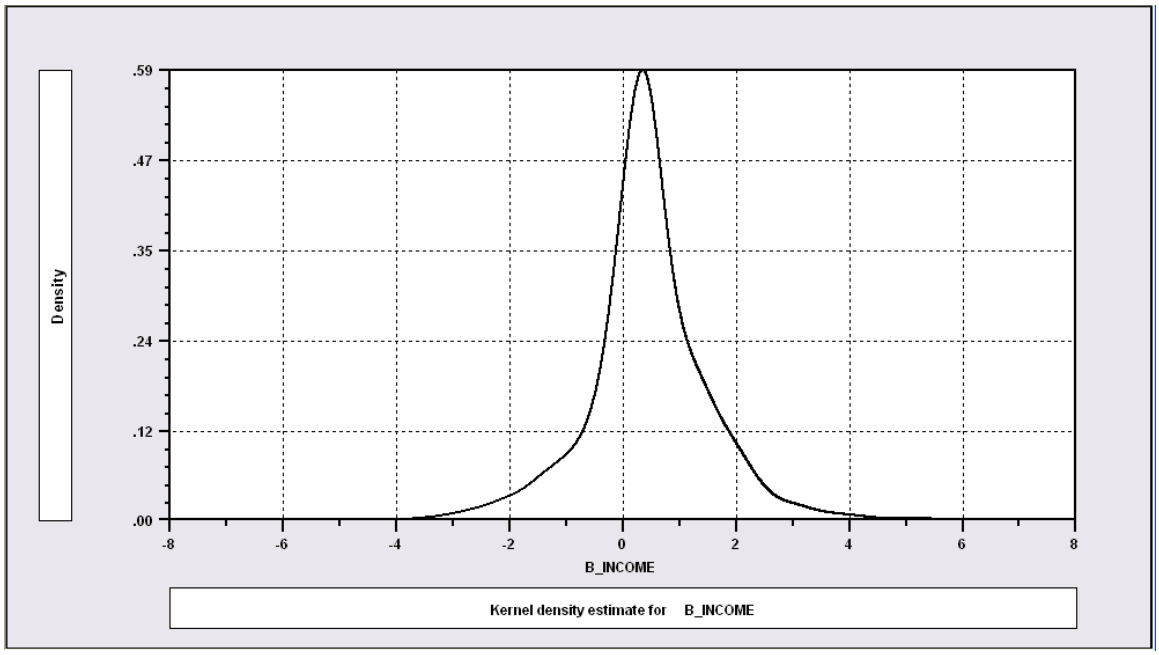


Figure 8.1 Kernel Density for Estimate of the Distribution of Means of Income Coefficient

8.2 Latent Class and Finite Mixture Modeling

Latent class modeling [see McLachlan and Peel (2000)] provides an alternative approach to accommodating heterogeneity. [Applications include Everitt (1988) and Uebersax (1999).] The natural approach assumes that parameter vectors, β_i are distributed among individuals with a discrete distribution, rather than the continuous distribution of the previous section. Thus, it is assumed that the population consists of a finite number, Q, of groups of individuals. The groups are heterogeneous, with common parameters, $\gamma_q = (\beta_q, \mu_q)$ for the members of the group, but the groups themselves are different from one another. The analyst does not know from the data which observation is in which class. (Hence the term *latent* classes.)

The model assumes that individuals are distributed heterogeneously with a discrete distribution in a population. Two other interpretations of the model are useful. The latent class model can also be viewed as a discrete approximation to the continuous distribution. This follows the development of Heckman and Singer (1984) who used this approach to modeling heterogeneity in a study of duration [see Section 12.6.1]. Alternatively, the *finite mixture* model may be used as a technique to model the distribution in its own right. This technique is often used to mix normal distributions to obtain a non-normal mixture distribution.

8.2.1 The Latent Class Ordered Choice Model

Class membership is distributed with discrete distribution,

$$\text{Prob(individual } i \text{ is a member of class} = q) = \pi_{iq} = \pi_q. \tag{8.15}$$

This statement needs its own interpretation. It can be given a long run frequency interpretation in that the probability that an individual drawn at random from the full population is a member of the particular class. Alternatively, it reflects the priors of the analyst over the same random outcome. Under either interpretation, then

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = \Sigma_q \, \text{Prob}(y_i = j \mid \mathbf{x}_i, \text{class} = q)\text{Prob}(\text{class} = q). \tag{8.16}$$

Combining terms from earlier, then, a latent class ordered probit model would be

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = \sum_{q=1}^{Q} \pi_q \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) \right]. \tag{8.17}$$

(We will use the probit formulation for this discussion. A logit model is obtained trivially by changing the assumed cdf and density – it will be a simple change of notation.) By this construction, the implied estimator of the cell probabilities would be a mixture of the class specific probabilities, using the estimated class probabilities, π_q for the mixture. Likewise, the partial effects would be

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \sum_{q=1}^{Q} \pi_q \left[\phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) - \phi\left(\mu_{j,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) \right] \boldsymbol{\beta}_q, \tag{8.18}$$

that is, the same weighted mixture of the class specific partial effects.

8.2.2 Estimation by Maximum Likelihood

The estimation problem now includes estimation of $(\boldsymbol{\beta}_q, \boldsymbol{\mu}_q, \pi_q), q = 1,...,Q$. The class probabilities are estimated with the other parameters. It is necessary to force the class probabilities to be between zero and one and to sum to one. A convenient way to do so is to use a multinomial logit parameterization of the class probabilities.

$$\pi_q = \frac{\exp(\theta_q)}{\Sigma_{q=1}^{Q} \exp(\theta_q)}, \; q = 1,...,Q, \; \theta_Q = 0. \tag{8.19}$$

Assembling the parts, then, the full log likelihood for the parameters, given the observed data is

$$\log L = \sum_{i=1}^{N} \log \left\{ \sum_{q=1}^{Q} \frac{\exp(\theta_q)}{\Sigma_{s=1}^{Q} \exp(\theta_s)} \sum_{j=0}^{J} m_{ij} \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q' \mathbf{x}_i\right) \right] \right\}, \tag{8.20}$$

where

$m_{ij} = 1$ if $y_i = j$ and 0 otherwise, $j = 0,...,J$; $i = 1,...,N$,

and the full vector of parameters to be estimated is

$$\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\mu}_1, ..., \boldsymbol{\beta}_Q, \boldsymbol{\mu}_Q, \theta_1, ..., \theta_Q),$$

with several constraints, $\mu_{-1,q} = -\infty$, $\mu_{0,q} = 0$, $\mu_{J,q} = +\infty$, $q = 1,...,Q$ and $\theta_Q = 0$.

We have assumed to this point that the number of classes, Q, is known. This will rarely be the case, so a question naturally arises, how can the analyst determine Q? Since Q is not a free parameter, a likelihood ratio test is not appropriate, though, in fact, log$L$ will increase when Q increases. Researchers typically use an information criterion, such as AIC, to guide them toward the appropriate value. [See Bhat (1996a) and Section 12.6.1.] Heckman and Singer (1984) note a practical guidepost. If the model is fit with too many classes, then estimates will become imprecise, even varying wildly. Signature features of a model that has been overfit will be exceedingly small estimates of the class probabilities (see below), wild values of the structural parameters and huge estimated standard errors. An application that illustrates this possibility appears in Section 2.11.

Statistical inference about the parameters can be done in the familiar fashion. The Wald test or likelihood ratio tests will probably be more convenient. Hypothesis tests across classes are unlikely to be meaningful. For example, suppose we fit a three class model. Tests about the equality of some of the coefficients in one class to those in another would probably be ambiguous, because the classes, themselves are indeterminate. It is rare that one can even put a name on the classes, other than, "1," "2," etc. Likewise, testing about the number of classes is an uncertain exercise. Consider our two class example below. If the parameters of the two classes are identical, it would seem that there is a single class. The number of restrictions would seem to be the number of model parameters. However, there remain two class probabilities, π_1 and π_2. If the parameter vectors are the same, then regardless of the values of π_1 and π_2, there is only one class. Thus, the degrees of freedom for this test are ambiguous. The same log likelihood will emerge for any pair of probabilities that sum to one.

The log likelihood can be maximized using conventional gradient methods. [See Econometric Software (2007).] An alternative method, the *EM* algorithm [Dempster, Laird and Rubin (1977)], is particularly well suited to latent class modeling. Though generally slower than gradient methods such as Broyden, Fletcher, Goldfarb and Shanno [see Greene (2008a)], the *EM* method does have the advantage of great stability.

The *EM* algorithm is most effective in estimating the parameters of "missing data models." In the model we are examining, the missing data are

d_{iq} = 1 if individual i is a member of class q and 0 if not.

If d_{iq} were observed, then the "complete data" log likelihood could be written

$$\begin{aligned}\log L \mid \mathbf{d} &= \sum_{i=1}^{N} \sum_{q=1}^{Q} d_{iq} \log \sum_{j=0}^{J} m_{ij} \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) \right] \\ &= \sum_{q=1}^{Q} \sum_{i=1}^{N} d_{iq} \log \sum_{j=0}^{J} m_{ij} \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) \right] \qquad (8.21) \\ &= \sum_{q=1}^{Q} \left\{ \sum_{i=1}^{N_q} \log \sum_{j=0}^{J} m_{ij} \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) \right] \right\}.\end{aligned}$$

(In the second line, we have only reversed the order of the summations.) That is, if d_{iq} were known, then we could partition the log likelihood into separate log likelihoods for the Q classes and maximize each one separately. Maximization of this log likelihood would be done by separating the observations into the Q known groups and estimating a separate ordered choice model for each group of N_q observations.

Since d_{iq} is not observed, we must maximize the earlier log likelihood instead. The *E* (expectation) step of the *EM* algorithm requires derivation of the expectation of log$L|\mathbf{d}$ given the observed data, $y_i, \mathbf{x}_i$, $i=1,...,N$ and the parameters of the class specific models, $\boldsymbol{\beta}_q$ and $\boldsymbol{\mu}_q$. This, in

turn requires deriving $E[d_{iq}|y_i,\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q]$. Unconditionally, $E[d_{iq}] = \pi_q$. However, there is more information in the sample. The conditional mean function, $E[d_{iq}|y_i,\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q]$ (which is the expectation conditioned on $m_{ij},\mathbf{x}_i$ and the parameters $\boldsymbol{\beta}_q,\boldsymbol{\mu}_q$) is found as follows: the joint density of y_i and d_{iq} is

$$\begin{aligned} P(y_i,d_{iq}|\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q) &= P(y_i|d_{iq},\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)P(d_{iq}) \\ &= \text{Prob}(y_i = j \mid \text{class} = q,\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q) \times \pi_q. \end{aligned}$$

Using Bayes Theorem,

$$\begin{aligned} P(d_{iq} \mid y_i,\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q) &= \frac{P(y_i \mid d_{iq},\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)P(d_{iq})}{P(y_i \mid \mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)} \\ &= \frac{P(y_i \mid d_{iq},\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)P(d_{iq})}{\Sigma_{q=1}^{Q}P(y_i \mid d_{iq},\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)P(d_{iq})} \\ &= \frac{\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q}{\sum_{q=1}^{Q}\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q}. \end{aligned}$$

The conditional mean is, then,

$$\begin{aligned} E(d_{iq} \mid y_i,\mathbf{x}_i,\boldsymbol{\beta}_q,\boldsymbol{\mu}_q) &= \sum_{q=1}^{Q} d_{iq}\frac{\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q}{\sum_{q=1}^{Q}\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q} \\ &= \frac{\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q}{\sum_{q=1}^{Q}\left\{\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}\pi_q} \\ &= \hat{w}_{iq}. \end{aligned}$$

(There is only one nonzero term in the summation in the first line.) The *M* (maximization) step of the EM algorithm consists of maximizing $E[\log L|\mathbf{d}]$ by replacing d_{iq} with the expectations derived above. (Note, we are conditioning on an existing (previous) value of $(\boldsymbol{\beta}_q,\boldsymbol{\mu}_q)$, so $\hat{w}_{iq}$ is not a function of the parameters in the expected log likelihood.) Thus, the *M* step consists of maximizing

$$\log L \mid E = \sum_{q=1}^{Q}\sum_{i=1}^{N}\hat{w}_{iq}\log\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]. \qquad (8.22)$$

Since the weights, $\hat{w}_{iq}$, are now known, this maximand can be partitioned into Q separate weighted log likelihoods that can be maximized separately;

$$\log L \mid E,q = \sum_{i=1}^{N}\hat{w}_{iq}\log\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right], q = 1,...,Q. \qquad (8.23)$$

To assemble the parts, then, the *EM* algorithm for latent class modeling as a general template that we can use for our ordered choice model is

(1) Obtain starting values for $\boldsymbol{\beta}_q, \boldsymbol{\mu}_q$, $q = 1,...,Q$.
(2) Compute weights $\hat{w}_{iq}$, $i = 1,...,N$ based on each of the q parameter vectors.
(3) Using the weights obtained in step (2), compute Q new sets of parameters by maximizing Q separate weighted log likelihoods.
(4) Return to step (2) if the new estimates are not sufficiently close to the previous ones. Otherwise, exit the iterations.

As noted earlier, this algorithm can take many iterations. However, each iteration is simple. Adding weights to the log likelihood we have been manipulating all along, is a trivial modification. Moreover, as shown by Dempster et al. (1977), the log likelihood increases with every iteration – that is the stability aspect that is not necessarily achieved by other gradient methods.

We note a few practical points: (1) It would be tempting to obtain the starting values by using, for each class, the single class estimates obtained by maximizing the log likelihood for the sample without the latent class structure. Unfortunately, this leads to a frustrating result. If the parameters in the classes are the same, then the sets of weights for the classes will also be the same, which means that the next set of parameter estimates will again be the same. The end result is that these starting values will prevent the iterations from ever reaching the solution. A practical expedient is a small, different perturbation of the original estimates for each class. (2) The EM algorithm finds the maximizer of the log likelihood function, but unlike other gradient methods, it does not automatically produce an estimate of the asymptotic covariance matrix of the estimator. That must be obtained separately after the estimation is done. Note that the second derivatives matrix (or an approximation to it) computed from the weighted log likelihood function is not an appropriate estimator of the asymptotic covariance matrix of the class specific parameter vector. (3) To this point, we have not obtained an estimator of π_q. The appropriate estimator, perhaps not surprisingly, is

$$\hat{\pi}_q = \frac{\Sigma_{i=1}^{N} \hat{w}_{iq}}{N} = \bar{\hat{w}}_q \,. \tag{8.24}$$

8.2.4 Estimating the Class Assignments

There is a secondary estimation problem in the latent class setting, known as the "classification problem." Ex post, it would be useful to be able to assign observations to classes. If we could do this, then the classes would not be latent, and the model would be superfluous. However, one's best guess of the class from which observation i is drawn would be based on the posterior,

$$\text{Prob(individual } i \text{ is in class } q | y_i, \mathbf{x}_i, \boldsymbol{\beta}_q, \boldsymbol{\mu}_q) \;=\; \hat{w}_{iq} \,,$$

as computed earlier. Thus, the EM algorithm provides the sample estimator for the classification problem automatically. If the EM algorithm has not been used, it is still possible to compute $\hat{w}_{iq}$ using the estimated parameters, simply using the definition given earlier. The end result would be to estimate the class membership for individual i as that q associated with the maximum value of $\hat{w}_{iq}$ for $q = 1,...,Q$.

8.2.5 A Latent Class Model Extension

The latent class interpretation of the model suggests a useful estension of the class probabilities model. Thus far, the specification provides no prior information about the class membership. That is, the prior class probabilities are constants,

$$\text{Prob}(\text{class} = q) = \pi_q.$$

If there were useful, though not definitive, sample information (in which case, the classes would not be latent) for determining class membership, then we might write

$$\text{Prob}(\text{class} = q \mid \mathbf{z}_i) = \pi_q(\mathbf{z}_i).$$

where presumably, $\mathbf{z}_i$ does not appear in the main model. For example, in our ordered probit model, it might be suspected that gender or working status has an influence on the class probabilities for health satisfaction. This is straightforward to build into the multinomial logit model, in the form

$$\pi_{iq} = \frac{\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}{\Sigma_{q=1}^{Q}\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z})},\ q = 1,...,Q,\ \theta_Q = 0,\ \boldsymbol{\delta}_q = \mathbf{0}. \tag{8.25}$$

Estimation is also only slightly more complicated. The log likelihood for the full model would now be

$$\log L = \sum_{i=1}^{N}\log\left\{\sum_{q=1}^{Q}\ \frac{\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}{\Sigma_{q=1}^{Q}\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}\sum_{j=0}^{J} m_{ij}\left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_i\right)\right]\right\}. \tag{8.26}$$

The EM algorithm would require a slight modification. We would add a step (2a) that would be estimation of the logit parameters, $(\theta_q,\boldsymbol{\delta}_q)$,$q$=1,...,$Q$-1 (with $\theta_Q = 0$ and $\boldsymbol{\delta}_Q = \mathbf{0}$). This (sub)step is done by fitting a multinomial logit model to the weights, $\hat{w}_{iq}$ based on proportions, rather than individual data. The implied log likelihood function is

$$\begin{aligned}\log L(\boldsymbol{\theta},\boldsymbol{\Delta}) &= \sum_{i=1}^{N}\ \sum_{q=1}^{Q}\ \hat{w}_{iq}\log\left[\frac{\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}{\Sigma_{q=1}^{Q}\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}\right] \\ &= \sum_{i=1}^{N}\ \sum_{q=1}^{Q}\ \hat{w}_{iq}\log\Lambda_{iq}.\end{aligned} \tag{8.27}$$

The solution to this estimation is also straightforward using Newton's method. The first order conditions are revealing of the structure of the problem;

$$\frac{\partial\log L(\boldsymbol{\theta},\boldsymbol{\Delta})}{\partial\begin{pmatrix}\theta_q \\ \boldsymbol{\delta}_q\end{pmatrix}} = \sum_{i=1}^{N}\ \left(\hat{w}_{iq} - \Lambda_{iq}\right)\begin{pmatrix}1 \\ \mathbf{z}_i\end{pmatrix}. \tag{8.28}$$

The first of the equations would imply Σ_i $(\hat{w}_{iq}$ - $\Lambda_{iq})$. If there were no covariates, $\mathbf{z}_i$ in the equation, this would return the original solution for $\hat{\pi}_q$ that was shown earlier. Thus, we find (as

Table 8.4 Estimated Two Class Latent Class Ordered Probit Model

Latent Class / Ordered Probit Model									
Number of observations	4483								
Log likelihood function	-5716.627	-5683.202							
Info. Criterion: AIC =	2.55883	2.54526							
Variable	Estimate	S.E.	b/s.e.	Prob	Estimate	S.E.	b/s.e.	Prob	OrdPrbt.
Parameters for Latent Class 1									
Constant	2.9502	.4131	7.142	.0000	2.6740	.9877	2.707	.0068	1.9788
AGE	-.0112	.0036	-3.302	.0010	-.0168	.0031	-5.491	.0000	-.0181
EDUC	.0066	.0200	.330	.7415	.0565	.0141	4.018	.0001	.0356
INCOME	-.8932	.3211	-2.782	.0054	-.0722	.2054	-.352	.7251	.2587
MARRIED	-.0038	.0841	-.046	.9635	-.1250	.0714	-1.751	.0800	-.0310
KIDS	-.0601	.0859	-.699	.4844	.0585	.0695	.841	.4001	.0606
MU(1)	1.0594	.2089	5.072	.0000	1.2427	.7287	1.705	.0881	1.1483
MU(2)	2.9914	.2411	12.406	.0000	3.1004	.9753	3.179	.0015	2.5478
MU(3)	3.2639	.1974	16.535	.0000	3.8124	1.045	3.648	.0003	3.0564
Parameters for Latent Class 2									
Constant	1.3384	.3151	4.247	.0000	1.7882	.3586	4.987	.0000	
AGE	-.0314	.0049	-6.403	.0000	-.0222	.0050	-4.428	.0000	
EDUC	.0760	.0214	3.551	.0004	.0063	.0298	.210	.8335	
INCOME	1.8767	.4844	3.874	.0001	.4473	.3481	1.285	.1988	
MARRIED	-.1106	.0962	-1.150	.2503	-.0611	.1142	-.535	.5924	
KIDS	.2182	.1014	2.151	.0315	.1243	.1110	1.120	.2627	
MU(1)	1.5400	.1661	9.273	.0000	1.4529	.3011	4.826	.0000	
MU(2)	2.4763	.1642	15.085	.0000	2.3938	.3921	6.105	.0000	
MU(3)	3.7191	.3423	10.865	.0000	2.3938	.3077	7.781	.0000	
+	Multinomial Logit Model for Class Probabilities								
ONE_1					.7383	.7621	.969	.3327	
FEMALE_1					-.0431	.1278	-.337	.7362	
HANDDU_1					-1.223	.2389	-5.120	.0000	
WORKIN_1					.4096	.1512	2.710	.0067	
ONE_2					.0000	..(Fixed Parameter)			
FEMALE_2					.0000	..(Fixed Parameter)			
HANDDU_2					.0000	..(Fixed Parameter)			
WORKIN_2					.0000	..(Fixed Parameter)			
+	Prior probabilities for class membership								
Class 1	.57532				.87182				1.00000
Class 2	.42468				.29818				0.00000

Table 8.5 Estimated Partial Effects from Latent Class Models

Summary of Marginal Effects for Ordered Probability Models										
Effects computed at means. Effects for binary variables are computed as differences of probabilities, other variables at means.										
	Latent Class Model					Expanded Latent Class Model				
Outcome	AGE	EDUC	INCOME	MARRIED	KIDS	AGE	EDUC	INCOME	MARRIED	KIDS
Y = 00	.0014	-.0024	-.0192	.0033	-.0039	.0092	-.0033	.0058	.0069	-.0030
Y = 01	.0053	-.0094	-.0741	.0128	-.0152	.0020	-.0073	.0128	.0157	-.0069
Y = 02	-.0012	.0022	.0770	-.0027	.0334	.0013	-.0047	.0082	.0109	-.0043
Y = 03	-.0031	.0055	.0172	-.0076	.0089	-.0012	.0045	-.0078	-.0094	.0041
Y = 04	-.0023	.0042	.0327	-.0058	.0068	-.0030	.0109	-.0190	-.0240	.0100

expected) that the ordinary methods and the EM method find the same maximizer of the log likelihood.

8.2.6 Application

Table 8.4 presents estimates of a two class latent class model using our base specification. The single class estimates are presented for comparison. The estimates for the two class model, as expected, bracket the one class estimates. Although the log likelihood has increased substantially (from -5752.985 to -5716.627), the class definition does not appear to have greatly changed the results. The estimated prior class probabilities are near 50%. In Table 8.4, we have also listed estimates of an extended model in which gender (*FEMALE*), handicapped (*HANDDUM*) and work status (*WORKING*) enter the class probabilities. This modification does appear to add significantly to the class segregation. Evidently *HANDDUM* and *WORKING*, though not *FEMALE*, are significant determinants. The log likelihood for the extended model jumps to -5683.202. The chi squared for the extension is 2(5716.627 – 5683.202) = 66.85 with 3 degrees of freedom, which is also highly significant. Partial effects for the two models are shown in Table 8.5.

8.2.7 Endogenous Class Assignment and A Generalized Ordered Choice Model

Greene, Harris, Hollingsworth and Maitra (2008) analyzed obesity in a sample of 12,601 men and 15,259 women in the U.S. National Health Interview Survey from 2005. The central feature of their model is a three outcome ordered choice model for weight class defined as normal, overweight and obese. Obesity is measured by the World Health Organization's standard body mass index, or *BMI*. *BMI* is computed as the weight in Kg divided the square of the height in meters. Values under 18.5 are classified by WHO as underweight. The 2% of their sample in this class was deleted. The remaining three classes are normal (18.5,25], overweight (25,30] and obese, (30,∞). There are great differences across individuals in body fat and conditioning, and the *BMI* classification is at best only a loose categorization of the desired health level indicated. The authors reasoned that the latent regression model *with known thresholds* that might seem superficially to apply,

$$
\begin{aligned}
BMI_i^* &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim N[0,\sigma^2],\\
BMI_i &= 0 \text{ if } BMI_i^* \le 25,\\
&\quad\ 1 \text{ if } 25 < BMI_i^* \le 30,\\
&\quad\ 2 \text{ if } BMI_i^* > 30,
\end{aligned}
$$

would be too narrow, and would neglect several sources of heterogeneity. They opted instead for an ordered "choice" model, defined as

$$
\begin{aligned}
BMI_i^* &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \varepsilon_i \sim N[0,1],\\
WT_i &= 0 \text{ if } BMI_i^* \le 0,\\
&\quad\ 1 \text{ if } 0 < BMI_i^* \le \mu,\\
&\quad\ 2 \text{ if } BMI_i^* > \mu.
\end{aligned}
$$

A recent study in *Science* [Herbert, Gerry and McQueen (2006)] suggests that an obesity predisposing genotype is present in 10% of individuals. In the sample, roughly 25% of the sample is categorized as obese. This suggests that a latent class model might be appropriate and that the class division depends on more than just this (unobserved) geno-type. The study used a two class model, with

$$
\begin{aligned}
class_i^* &= \boldsymbol{\alpha}'\mathbf{w}_i + u_i,\ u_i \sim N[0,1],\\
class_i &= 0 \text{ if } class_i^* \leq 0,\\
&\quad\ 1 \text{ if } class_i^* > 0.
\end{aligned}
$$

The rigidity of the BMI classification, itself, might have produced erroneous classifications. For examples, athletes with high *BMI* levels due to high percentages of muscle mass, rather than fat, could be misclassified. To accommodate this sort of heterogeneity, the authors specified a heterogeneous threshold model,

$$\mu_i = \exp(\theta + \boldsymbol{\delta}'\mathbf{r}_i).$$

Finally, reasoning that the *BMI* outcome and the latent class assignment would likely depend on common features, both observed (in $\mathbf{x}_i$ and $\mathbf{w}_i$) and unobserved (in ε and u), they specified a joint normal distribution for ε_i and u_i with correlation ρ. (This is the first application of this model extension that we have seen. Since this is a cross section analysis, the natural extension is straightforward to build into the specification. If the sample were a panel, it would make sense to build a time invariant random effect into the main equation and allow that to be correlated with u_i in the class assignment. Some more elaborate specification would be necessary of the model specified more than two classes.

Combining all of the components, we have

Outcome Model:

$$
\begin{aligned}
(BMI_i^*|class = c) &= \boldsymbol{\beta}_c'\mathbf{x}_i + \varepsilon_{ic},\ \varepsilon_{ic} \sim N[0,1],\\
WT_i|class=c &= 0 \text{ if } BMI_i^*|class = c \leq 0,\\
&\quad\ 1 \text{ if } 0 < BMI_i^*|class = c \leq \mu_{ic},\\
&\quad\ 2 \text{ if } BMI_i^*|class = c > \mu_{ic},\\
\text{Threshold}_i|\text{class}=c: &\ \mu_{ic} = \exp(\theta_c + \boldsymbol{\delta}_c'\mathbf{r}_i).
\end{aligned}
$$

Class Assignment:

$$
\begin{aligned}
c_i^* &= \boldsymbol{\alpha}'\mathbf{w}_i + u_i,\ u_i \sim N[0,1],\\
c_i &= 0 \text{ if } c_i^* \leq 0,\\
&\quad\ 1 \text{ if } c_i^* > 0.,
\end{aligned}
$$

Endogenous Class Assignment:

$$(\varepsilon_{ic}, u_i) \sim N_2[(0,0),(1,\rho_c,1)].$$

Formation of the probabilities for the observed outcomes is a bit more complicated than previously due to the correlation between the class assignment and the *BMI* outcome. Generically,

$$\text{Prob}[WT_i = j \mid \text{class} = c] = \text{Prob}[WT_i = j, class = \text{c}] / \text{Prob}(class = \text{c}).$$

To form the likelihood, we require the joint probabilites, not the conditional;

$$\log L = \sum_{i=1}^{N} \log\left[\sum_{class=0}^{1} \text{Prob}(class = c)\text{Prob}(WT_i = j \mid class = c)\right].$$

The joint probability is a bivariate normal probability. (See Section 9.1.) To reach the components of the log likelihood and the probabilities to analyze for the partial effects, we begin with

$$\text{Prob}(WT_i = j \mid class = c) = \frac{\left\{\begin{array}{l}\Phi_2[(\mu_{i,j,c} \quad -\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]- \\ \Phi_2[(\mu_{i,j-1,c}-\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]\end{array}\right\}}{\Phi[(2c-1)\boldsymbol{\alpha}'\mathbf{z}_i]}.$$

The unconditional probability is

$$\text{Prob}(WT_i = j) = \sum_{c=0}^{1} \text{Prob}(class = c)\frac{\left\{\begin{array}{l}\Phi_2[(\mu_{i,j,c} \quad -\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]- \\ \Phi_2[(\mu_{i,j-1,c}-\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]\end{array}\right\}}{\Phi[(2c-1)\boldsymbol{\alpha}'\mathbf{z}_i]},$$

$$=\sum_{c=0}^{1} \Phi[(2c-1)\boldsymbol{\alpha}'\mathbf{z}_i]\frac{\left\{\begin{array}{l}\Phi_2[(\mu_{i,j,c} \quad -\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]- \\ \Phi_2[(\mu_{i,j-1,c}-\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]\end{array}\right\}}{\Phi[(2c-1)\boldsymbol{\alpha}'\mathbf{z}_i]},$$

$$=\sum_{c=0}^{1} \left\{\begin{array}{l}\Phi_2[(\mu_{i,j,c} \quad -\boldsymbol{\beta}_c'\mathbf{x}),((2c-1)\boldsymbol{\alpha}'\mathbf{z}),((2c-1)\rho_c)]- \\ \Phi_2[(\mu_{i,j-1,c}-\boldsymbol{\beta}_c'\mathbf{x}),((2c-1)\boldsymbol{\alpha}'\mathbf{z}),((2c-1)\rho_c)]\end{array}\right\}.$$

Combining all terms, then,

$$\log L = \sum_{i=1}^{N} \log \sum_{c=0}^{1} \sum_{j=0}^{2} m_{ij} \left\{\begin{array}{l}\Phi_2[(\mu_{i,j,c}-\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]- \\ \Phi_2[(\mu_{i,j-1,c}-\boldsymbol{\beta}_c'\mathbf{x}_i),((2c-1)\boldsymbol{\alpha}'\mathbf{z}_i),((2c-1)\rho_c)]\end{array}\right\},$$

where $m_{ij} = j$ if $WT_i = j$, $j = 0,1,2$, $\mu_{i,-1,c} = -\infty$, $\mu_{i,0,c} = 0$, $\mu_{i,1,c} = \exp(\theta_c + \boldsymbol{\delta}_c'\mathbf{r}_i)$, $\mu_{i,2,c} = +\infty$.

In order to simplify the derivation of the partial effects, assume for the present that $\mathbf{x}_i$, $\mathbf{z}_i$ and $\mathbf{r}_i$ all contain the same variables, labeled $\mathbf{w}_i$. The partial effects will contain three terms, for the latent regression, the class assignment and the threshold model. For variables that appear in more than one part, the partial effect will be obtained by adding the terms. For convenience, we will drop the observation subscript. Partial effects will typically be computed at the means of the variables, or by averaging the partial effects over all observations Define the quantities

$$\begin{aligned} A_{jic} &= \mu_{j,c} - \boldsymbol{\beta}_c'\mathbf{w}_i, \\ B_{ic} &= (2c\text{-}1)\boldsymbol{\alpha}'\mathbf{w}_i, \\ \tau_c &= (2c\text{-}1)\rho_c. \end{aligned}$$

Then,

$$\text{Prob}(WT_i = j \mid \mathbf{w}) = \sum_{c=0}^{1} \Phi_2[A_{j,ic}, B_{ic}, \tau_c] - \Phi_2[A_{j-1,ic}, B_{ic}, \tau_c].$$

The partial effects are

$$\frac{\partial \text{Prob}(WT = j \mid \mathbf{x}_i)}{\partial \mathbf{w}_i} = \sum_{c=0}^{1} \begin{bmatrix} \Phi\left(\frac{B_{ic} - \tau_c A_{j,ic}}{\sqrt{1-\tau_c^2}}\right)\left[\phi(A_{j,ic}) - \phi(A_{j-1,ic})\right](-\boldsymbol{\beta}_c) + \\ \Phi\left(\frac{B_{ic} - \tau_c A_{j,ic}}{\sqrt{1-\tau_c^2}}\right)\left[\phi(A_{j,ic})\mu_{j,c} - \phi(A_{j-1,ic})\mu_{j-1,c}\right](\boldsymbol{\delta}_c) + \\ \phi(B_c)\left[\Phi\left(\frac{A_{j,ic} - \tau_c B_{ic}}{\sqrt{1-\tau_c^2}}\right) - \Phi\left(\frac{A_{j-1,ic} - \tau_c B_{ic}}{\sqrt{1-\tau_c^2}}\right)\right](2c-1)(\boldsymbol{\alpha}_i) \end{bmatrix}.$$

8.3 Generalized Ordered Choice Model with Random Thresholds (3)

Cunha, Heckman and Navarro (2007) present a variety of settings in which an optimizing economic agent chooses an outcome along a continuum by revealing the discrete choice of an interval along the support of the underlying random variable. This would be consistent with our latent regression interpretation, though these authors present a much more abstract, economic theory based motivation. They cite a variety of applications including when to fell a tree and how many years of education to accomplish as examples, and argue that the general class of models is consistent with an ordered choice model. Some of these suggest an ordered choice model with random thresholds, an idea first suggested by Cameron and Heckman (1998) and later developed by Carniero, Hansen and Heckman (2001, 2003). The theoretical underpinnings of an ordered choice model with random thresholds are further developed by Heckman and Navarro (2007) and Vytlacil (2006).

In this section, we combine the features of the preceding generalized models in a single internally consistent model framework. The model contains random parameters, heterogeneous thresholds and heteroscedasticity. We depart from the base case,

$$\text{Prob}[y_i = j \mid \mathbf{x}_i] = F(\mu_{i,j} - \boldsymbol{\beta}_i'\mathbf{x}_i) - F(\mu_{i,j-1} - \boldsymbol{\beta}_i'\mathbf{x}_i) > 0, j = 0,1,...,J. \quad (8.29)$$

The intrinsic heterogeneity across individuals is captured by writing

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{v}_i, \quad (8.30)$$

where $\boldsymbol{\Gamma}$ is a lower triangular matrix and $\mathbf{v}_i \sim N[\mathbf{0},\mathbf{I}]$. Thus, $\boldsymbol{\beta}_i$ is normally distributed across individuals with conditional mean,

$$E[\boldsymbol{\beta}_i|\mathbf{x}_i,\mathbf{z}_i] = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i,$$

and conditional variance,

$$\text{Var}[\boldsymbol{\beta}_i|\mathbf{x}_i,\mathbf{z}_i] = \boldsymbol{\Gamma}\mathbf{I}\boldsymbol{\Gamma}' = \boldsymbol{\Omega}.$$

This is a random parameters formulation that appears elsewhere, e.g., Greene (2002, 2005) and Jones and Hensher (2004). It is the same as the random parameters model developed in the previous section.

The thresholds are modeled as

$$\mu_{i,j} = \mu_{i,j-1} + \exp(\alpha_j + \boldsymbol{\delta}'\mathbf{r}_i + \sigma_j w_{ij}), \mu_0 = 0, \mu_{-1} = -\infty, \mu_J = +\infty, w_{ij} \sim N[0,1]. \quad (8.31)$$

Integrating the difference equation, we obtain

$$\begin{aligned}
\mu_{i,1} &= \exp(\alpha_1 + \boldsymbol{\delta}'\mathbf{r}_i + \sigma_1 w_{j1}) \\
&= \exp(\boldsymbol{\delta}'\mathbf{r}_i)\exp(\alpha_1 + \sigma_1 w_{j1}), \\
\mu_{i,2} &= \exp(\boldsymbol{\delta}'\mathbf{r}_i)[\exp(\alpha_1 + \sigma_1 w_{j1}) + \exp(\alpha_2 + \sigma_2 w_{j2})], \\
\mu_{i,j} &= \exp(\boldsymbol{\delta}'\mathbf{r}_i)\left(\Sigma_{m=1}^{j}\exp(\alpha_m + \sigma_m w_{im})\right), \\
\mu_{i,J} &= +\infty \text{ is imposed by } \alpha_J = +\infty \text{ and } \sigma_J = 0.
\end{aligned} \tag{8.32}$$

An extension of the model along the lines suggested by King et al. (2004) for the vignettes model in Section 7.3 would be to allow each threshold parameter to have its own parameter vector. Then,

$$\mu_{i,j} = \mu_{i,j-1} + \exp(\alpha_j + \boldsymbol{\delta}_j'\mathbf{r}_i + \sigma_j w_{ij}),\ \mu_0 = 0,\ \mu_{-1} = -\infty,\ \mu_J = +\infty,\ w_{ij} \sim \mathrm{N}[0,1]. \tag{8.33}$$

So

$$\mu_{i,j} = \exp(\boldsymbol{\delta}'\mathbf{r}_i)\sum_{m=1}^{j}\exp(\alpha_m + \boldsymbol{\delta}_m'\mathbf{r}_i + \sigma_m w_{im}).$$

For simplicity in what follows, we will maintain the simpler model with a common slope vector in the thresholds.

This model preserves the ordering of the thresholds and incorporates the necessary normalizations. Note that the thresholds, like the regression itself, are shifted by both observable ($\mathbf{r}_i$) and unobservable (w_{ij}) heterogeneity. The model is fully consistent in that probabilities are all positive and sum to one by construction. Finally, the disturbance variance is allowed to be heteroscedastic, as before, randomly as well as deterministically; thus,

$$\mathrm{Var}[\varepsilon_i|\mathbf{h}_i] = [\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_i)]^2, \tag{8.34}$$

where $e_i \sim \mathrm{N}[0,1]$.

Let $\mathbf{v}_i = (v_{i1},...,v_{iK})'$ and $\mathbf{w}_i = (w_{i1},...,w_{i,J-1})'$. Combining terms, the conditional probability of outcome j is

$$\mathrm{Prob}[y_i = j \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{h}_i,\mathbf{r}_i,\mathbf{v}_i,\mathbf{w}_i,e_i] = F\left[\frac{\mu_{ij} - \boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_i)}\right] - F\left[\frac{\mu_{i,j-1} - \boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_i)}\right]. \tag{8.35}$$

The term that enters the log likelihood function is unconditioned on the unobservables. Thus, after integrating out the unobservable heterogeneity, we have

$$\mathrm{Prob}[y_i = j \mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{h}_i,\mathbf{r}_i] = \int_{\mathbf{v}_i,\mathbf{w}_i,e_i}\left(F\left[\frac{\mu_{ij} - \boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_i)}\right] - F\left[\frac{\mu_{i,j-1} - \boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_i)}\right]\right) f(\mathbf{v}_i,\mathbf{w}_i,e_i)\,d\mathbf{v}_i\,d\mathbf{w}_i\,de_i. \tag{8.36}$$

where

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \mathbf{L}\mathbf{D}\mathbf{v}_i,$$

and

$$\mu_{ij} = \exp(\boldsymbol{\delta}'\mathbf{r}_i)\left(\Sigma_{m=1}^{j}\exp(\alpha_m + \sigma_m w_{im})\right), j = 1,...,J\text{-}1.$$

The model is estimated by maximum simulated likelihood. The simulated log likelihood function is

$$\log L_S(\boldsymbol{\beta},\boldsymbol{\Delta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{L},\mathbf{D},\boldsymbol{\sigma},\tau)=$$
$$\sum_{i=1}^{n} \log \frac{1}{R}\sum_{r=1}^{R} \left(F\left[\frac{\mu_{ij,r}-\boldsymbol{\beta}_{i,r}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_{i,r})}\right]-F\left[\frac{\mu_{i,j-1,r}-\boldsymbol{\beta}_{i,r}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_{i,r})}\right]\right). \quad (8.37)$$

This is the model in its full generality. Whether a particular data set is rich enough to support this much parameterization, particularly the elements of the covariances of the unobservables in $\boldsymbol{\Gamma}$, is an empirical question that will depend on the application. In estimation of a very similar model, Eluru, Bhat and Hensher (2008) found that a large number of zero restrictions on the various parameters was necessary to estimate the model. The extended model in (8.33) will likewise require a rich data set.

The model contains three points at which changes in the observed variables can induce changes in the probabilities of the outcomes, in the thresholds, in the utility function, and in the variance. The, probability of interest is

$$\text{Prob}(y_i=j\mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{h}_i,\mathbf{r}_i)=\int_{\mathbf{v}_i,\mathbf{w}_i,e_i}\left(\begin{array}{l}F\left[\dfrac{\mu_{ij}-(\boldsymbol{\beta}+\boldsymbol{\Delta}\mathbf{z}_i+\mathbf{LD}\mathbf{v}_i)'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]-\\ F\left[\dfrac{\mu_{i,j-1}-(\boldsymbol{\beta}+\boldsymbol{\Delta}\mathbf{z}_i+\mathbf{LD}\mathbf{v}_i)'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]\end{array}\right) f(\mathbf{v}_i,\mathbf{w}_i,e_i)d\mathbf{v}_i d\mathbf{w}_i de_i,$$
$$\mu_{ij} = \exp(\boldsymbol{\delta}'\mathbf{r}_i)\left(\Sigma_{m=1}^{j}\exp(\alpha_m+\sigma_m w_{im})\right), j=1,\ldots,J\text{-}1. \quad (8.38)$$

The set of partial effects is

$$\frac{\partial \text{Prob}(y_i=j\mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{h}_i,\mathbf{r}_i)}{\partial \mathbf{x}_i}=$$
$$\int_{\mathbf{v}_i,\mathbf{w}_i,e_i}\left(\frac{1}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\left\{\begin{array}{l}f\left[\dfrac{\mu_{ij}-\boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]-\\ f\left[\dfrac{\mu_{i,j-1}-\boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]\end{array}\right\}\left(-(\boldsymbol{\beta}+\boldsymbol{\Delta}\mathbf{z}_i+\mathbf{LD}\mathbf{v}_i)\right)\right) f(\mathbf{v}_i,\mathbf{w}_i,e_i)d\mathbf{v}_i d\mathbf{w}_i de_i$$
(8.39a)

$$\frac{\partial \text{Prob}(y_i=j\mid \mathbf{x}_i,\mathbf{z}_i,\mathbf{h}_i,\mathbf{r}_i)}{\partial \mathbf{z}_i}=$$
$$\int_{\mathbf{v}_i,\mathbf{w}_i,e_i}\left(\frac{1}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\left\{\begin{array}{l}f\left[\dfrac{\mu_{ij}-\boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]-\\ f\left[\dfrac{\mu_{i,j-1}-\boldsymbol{\beta}_i'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{h}_i+\tau e_i)}\right]\end{array}\right\}\left(-\boldsymbol{\Delta}'\mathbf{x}_i\right)\right) f(\mathbf{v}_i,\mathbf{w}_i,e_i)d\mathbf{v}_i d\mathbf{w}_i de_i. \quad (8.39b)$$

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, \mathbf{r}_i)}{\partial \mathbf{h}_i} =$$

$$\int_{\mathbf{v}_i, \mathbf{w}_i, e_i} \left(\left\{ f\left[\frac{\mu_{ij} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right] \left(\frac{\mu_{ij} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right) - f\left[\frac{\mu_{i,j-1} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right] \left(\frac{\mu_{i,j-1} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right) \right\} (-\boldsymbol{\gamma}) \right) f(\mathbf{v}_i, \mathbf{w}_i, e_i) d\mathbf{v}_i d\mathbf{w}_i de_i \tag{8.39c}$$

$$\frac{\partial \text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, \mathbf{r}_i)}{\partial \mathbf{r}_i} =$$

$$\int_{\mathbf{v}_i, \mathbf{w}_i, e_i} \left(\left\{ f\left[\frac{\mu_{ij} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right] \left(\frac{\mu_{ij}}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right) - f\left[\frac{\mu_{i,j-1} - \boldsymbol{\beta}_i' \mathbf{x}_i}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right] \left(\frac{\mu_{i,j-1}}{\exp(\boldsymbol{\gamma}' \mathbf{h}_i + \tau e_i)} \right) \right\} (\boldsymbol{\delta}) \right) f(\mathbf{v}_i, \mathbf{w}_i, e_i) d\mathbf{v}_i d\mathbf{w}_i de_i \tag{8.39d}$$

Effects for particular variables that appear in more than one part of the model are added from the corresponding parts.

Development of a generalized model along these lines appears in Eluru, Bhat and Hensher (2008). An application appears in Greene and Hensher (2009). The former is a study of extent of injuries in traffic accidents. In the latter, the authors examine the information processing strategies in commuter choices of travel routes. Table 8.6 below shows an application of the model (with some of its features) to our health satisfaction example.

Table 8.6 Estimated Generalized Random Thresholds Ordered Logit Model

Random Thresholds Ordered Choice Model				
Dependent variable		HEALTH		
Number of observations		4483		
Log likelihood function		-5725.181		
Info. Criterion: AIC =		2.56310		
Underlying probabilities based on Logistic				
Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]
Latent Regression Equation				
Constant	11.7009	1.4905	7.850	.0000
AGE	-.1330	.0205	-6.496	.0000
EDUC	.3236	.0667	4.853	.0000
INCOME	2.2877	.7782	2.940	.0033
MARRIED	-.3397	.3095	-1.097	.2724
KIDS	.6054	.3061	1.978	.0479
Intercept Terms in Random Thresholds				
Alpha-01	1.7060	.1429	11.936	.0000
Alpha-02	2.2777	.1571	14.501	.0000
Alpha-03	1.8926	4.8200	.393	.6946
Standard Devs. of Random Thresholds				
Alpha-01	.5195	.1721	3.019	.0025
Alpha-02	.1995	.0616	3.239	.0012
Alpha-03	4.2325	16.2463	.261	.7945
Standard Devs. of Random Regr. Params.				
Constant	2.5004	1.0499	2.382	.0172
AGE	.0407	.0135	3.027	.0025
EDUC	.0050	.0626	.080	.9362
INCOME	.6391	1.5347	.416	.6771
MARRIED	.5556	.3146	1.766	.0773
KIDS	.1233	.5756	.214	8304
Heteroscedasticity in Regr. Equation				
FEMALE	.0020	.0532	.038	.9698
Latent Heterogeneity in Var. of Eps.				
Tau(v)	.3073	.1503	2.045	.0409

9

Ordered Choice Modeling with Panel and Time Series Data

Development of models for panel data parallel those in other modeling settings. The departure point is the familiar fixed and random effects approaches. We then consider other types of applications including extensions of the random parameters and latent classes formulations, dynamic models and some special treatments that accommodate features peculiar to the ordered choice models.

9.1 Ordered Choice Models with Fixed Effects

An ordered choice model with fixed effects formulated in the most familiar fashion would be

$$\text{Prob}[y_{it} = j \mid \mathbf{x}_i] = F(\mu_j - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{it}) - F(\mu_{j-1} - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{it}) > 0, j = 0,1,...,J. \quad (9.1)$$

At the outset, there are two problems that this model shares with other nonlinear fixed effects models. First, regardless of how estimation and analysis are approached, time invariant variables are precluded. Since social science applications typically include demographic variables such as gender and, for some at least, education level, that are time invariant, this is likely to be a significant obstacle. (Several of the variables in the GSOEP analyzed by Boes and Winkelmann (2006b) and others are time invariant.) Second, there is no sufficient statistic available to condition the fixed effects out of the model. That would imply that in order to estimate the model as stated, one must maximize the full log likelihood,

$$\log L = \sum_{i=1}^{N} \log \left\{ \prod_{t=1}^{T_i} \left(\sum_{j=0}^{J} m_{ijt} \left[\Phi\left(\mu_j - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{it}\right) - \Phi\left(\mu_{j-1} - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{i}\right) \right] \right) \right\}. \quad (9.2)$$

If the sample is small enough, of course, one may simply insert the individual group dummy variables and treat the entire pooled sample as a cross section. See, e.g., Mora (2006) for a cross-country application in banking that includes separate country dummy variables. We are interested, instead, in the longitudinal data case in which this would not be feasible. The data set from which our sample used in the preceding examples is extracted comes from an unbalanced panel of 7,293 households, observed from 1 to 7 times each.

The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Greene (2004a,b and 2008a, Section 16.9.6.c). The likelihood function is globally concave [see Pratt (1981)], so despite its superficial complexity, the estimation is straightforward. In another application, based on the full panel data set [see Greene (2008a, pp. 838-840), estimation of the full model required roughly five seconds of computation on an ordinary desktop computer.

The larger methodological problem with this approach would be at least the potential for the incidental parameters problem that has been widely documented for the binary choice case. [See, e.g., Lancaster (2000).] That is the small T bias in the estimated parameters when the full MLE is applied in panel data. For $T = 2$ in the binary logit model, it has been shown analytically [Abrevaya (1997)] that the full MLE converges to $2\boldsymbol{\beta}$. [See, as well, Hsiao (1986, 2003).] No corresponding results have been obtained for larger T or for other models. However, Monte

Carlo results have strongly suggested that the small sample bias persists for larger T as well, though as might be expected, it diminishes with increasing T.

No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) result on the small T bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model. The Monte Carlo results in Greene (2004b) reproduced below in Figure 9.1 suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. (In the first, third and fifth rows that correspond to estimation of coefficients, the true coefficients being estimated both equal one.)

Table 2. Means of empirical sampling distributions, $N = 1{,}000$ individuals based on 200 replications.

	$T=2$		$T=3$		$T=5$		$T=8$		$T=10$		$T=20$	
	β	δ	β	δ	β	δ	β	δ	β	δ	β	δ
Logit Coeff	2.020	2.027	1.698	1.668	1.379	1.323	1.217	1.156	1.161	1.135	1.069	1.062
Logit M.E.[a]	1.676	1.660	1.523	1.477	1.319	1.254	1.191	1.128	1.140	1.111	1.034	1.052
Probit Coeff	2.083	1.938	1.821	1.777	1.589	1.407	1.328	1.243	1.247	1.169	1.108	1.068
Probit M.E[a].	1.474	1.388	1.392	1.354	1.406	1.231	1.241	1.152	1.190	1.110	1.088	1.047
Ord. Probit	2.328	2.605	1.592	1.806	1.305	1.415	1.166	1.220	1.131	1.158	1.058	1.068

[a]Average ratio of estimated marginal effect to true marginal effect.

Figure 9.1 Monte Carlo Analysis of Biases in Fixed Effects MLE in Discrete Choice Models

The preceding bode ill for unconditional fixed effects models for ordered choice. So far, the approach has little to recommend it other than the theoretical robustness of fixed effects as an alternative to random effects. Recent proposals for "bias reduction" estimators for binary choice models, including Fernandez-Val and Vella (2007), Fernandez-Val (2008), Carro (2007), Hahn and Newey (2004) and Hahn and Kuersteiner (2003) suggest some directions for further research. However, no counterparts for the ordered choice models have yet been developed. We would note, for this model, the estimation of $\boldsymbol{\beta}$ which is the focus of these estimators, is only a means to the end. As seen earlier, in order to make meaningful statements about the implications of the model for behavior, it will be necessary to compute probabilities and derivatives. These, in turn, will require estimation of the constants, or some surrogates. The problem remains to be solved.

In their application to the GSOEP panel data set, Boes and Winkelmann (2006b) further modify the heterogeneous thresholds model. Their model is a fixed effects model,

$$\text{Prob}[y_{it}=j \mid \mathbf{x}_{i1}, \mathbf{x}_{i2},...,\mathbf{x}_{iT}] = F(\mu_{ij} - \boldsymbol{\beta}_j'\mathbf{x}_{it}) - F(\mu_{i,j-1} - \boldsymbol{\beta}_{j-1}'\mathbf{x}_i), \tag{9.3}$$

where

$$\mu_{ij} = \mu_j + \alpha_i.$$

Seeking to avoid the incidental parameters problem, they use Mundlak's (1978) and Chamberlain's (1980) device to model the fixed effect. Projecting the fixed effects on the group means of the regressors,

$$\alpha_i = \boldsymbol{\gamma}_j'\bar{\mathbf{x}}_i + \sigma v_i, \tag{9.4}$$

they obtain an equivalent random effects model,

$$\text{Prob}[y_{it}=j \mid \mathbf{x}_{i1}, \mathbf{x}_{i2},...,\mathbf{x}_{iT}] = F(\mu_{ij} - \boldsymbol{\beta}_j'\mathbf{x}_{it}) - F(\mu_{i,j-1} - \boldsymbol{\beta}_{j-1}'\mathbf{x}_i),$$

Where

$$\mu_{ij} = \mu_j + \boldsymbol{\gamma}_j'\bar{\mathbf{x}}_i + \sigma v_i,\ v_i \sim \text{N}[0,1], \tag{9.5}$$

and σ is a new parameter to be estimated. This model is estimated by using quadrature to integrate v_i out of the log likelihood. [See the next section and Butler and Moffitt (1982) for the methodology.] As observed at several earlier points, the placement of the heterogeneity in the thresholds is not substantive; it can be moved to the mean of the regression with no change in the interpretation of the model. As usual, the placement of the fixed effects in this linear specification is not consequential. Thus, their model is functionally equivalent to a more conventional random effects model with the group means added as covariates;

$$\text{Prob}[y_{it}=j \mid \mathbf{x}_{i1,}\mathbf{x}_{i2},...,\mathbf{x}_{iT}] = F[\mu_j-(\boldsymbol{\beta}_j'\mathbf{x}_{it}+\boldsymbol{\gamma}_{*j}'\bar{\mathbf{x}}_i+\sigma v_i)]-F[\mu_{j\text{-}1}-(\boldsymbol{\beta}_{j\text{-}1}'\mathbf{x}_i+\boldsymbol{\gamma}_{*j-1}'\bar{\mathbf{x}}_i+\sigma v_i)]. \quad (9.6)$$

The underlying logic of the Brant test suggests an alternative approach to estimation proposed by Das and van Soest (2000). Consider the base case ordered logit model with fixed effects. The model assumptions imply that

$$\begin{aligned}\text{Prob}[y_{it}>j \mid \mathbf{x}_{it}] &= \Lambda(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{it}-\mu_j) \\ &= \Lambda[(\alpha_i-\mu_j)+\boldsymbol{\beta}'\mathbf{x}_{it}].\end{aligned} \quad (9.7)$$

Now, define a binary variable $w_{it,j} = 1[y_{it}>j], j = 0,1,\ldots,J\text{-}1$. It follows that

$$\begin{aligned}\text{Prob}[y_{it}>j \mid \mathbf{x}_{it}] &= \Lambda[(\alpha_i-\mu_j)+\boldsymbol{\beta}'\mathbf{x}_{it}] \\ &= \Lambda[\lambda_i+\boldsymbol{\beta}'\mathbf{x}_{it}] \\ &= \text{Prob}(w_{itj}=1 \mid \mathbf{x}_{it}).\end{aligned} \quad (9.8)$$

The "j" specific part of the constant is the same for all individuals so it is absorbed in λ_i. Thus, a fixed effects binary logit model applies to each of the $J-1$ binary random variables, $w_{it,j}$. The method of Rasch (1960), Andersen (1970) and Chamberlain (1980) can be applied to each of these binary choice models to obtain an estimator of $\boldsymbol{\beta}$ without having to estimate the constant terms. [See also Greene (2008a, pp. 800-806).] This provides $J-1$ estimators of the parameter vector $\boldsymbol{\beta}$ (but no estimator of the threshold parameters). The authors propose to reconcile these different estimators by using a minimum distance estimator of the common true $\boldsymbol{\beta}$. The minimum distance estimator at the second step is chosen to minimize

$$q=\sum_{l=1}^{J-1}\sum_{m=1}^{J-1}\left(\hat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}\right)'\left[\mathbf{V}^{-1}\right]_{lm}\left(\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}\right), \quad (9.9)$$

where $\left[\mathbf{V}^{-1}\right]_{lm}$ is the l,m block of the inverse of the $(J-1)K\times(J-1)K$ partitioned matrix, $\mathbf{V}$, that contains $Asy.Cov\left(\hat{\boldsymbol{\beta}}_l,\hat{\boldsymbol{\beta}}_m\right)$. The appropriate form of this matrix for a set of cross-section estimators is given in Brant (1990). Since Das and van Soest (2000) used the counterpart for Chamberlain's fixed effects estimator, this would be inappropriate. They used, instead, a counterpart to the BHHH estimator. The l,m block of $\mathbf{V}$ (before inversion) is computed using

$$\mathbf{V}_{lm}=\sum_{i=1}^{N}\left(\frac{\partial\log L_{i,l}}{\partial\boldsymbol{\beta}_l}\right)\left(\frac{\partial\log L_{i,m}}{\partial\boldsymbol{\beta}_m}\right), \quad (9.10)$$

where $\log L_{i,m}$ is the contribution of individual i to the log likelihood for β_l. The diagonal blocks of the matrix are the BHHH estimators for the asymptotic covariance matrices for the j specific estimators.

As in the binary choice case, the complication of the fixed effects model is the small T bias, not the computation. The Das and van Soest approach finesses this problem—their estimator is consistent—but at the cost of losing the information needed to compute partial effects or predicted probabilities.

Winkelmann and Winkelmann (1998) analyzed data on well being from the German Socioeconomic Panel (GSOEP). The central question under the analysis is "How satisfied are you at present with your life as a whole?" which was answered on a discrete scale from 0 to 10. (See Section 2.1 for discussion of the methodological aspects of this analysis.) The natural approach to the analysis would be an ordered choice – the authors were interested in the effect of unemployment on the response. A fixed effects ordered choice (logit) model is the starting point for the specification.. Since there is no sufficient statistic available to use to condition the fixed effects out of the log likelihood, and fitting the fixed effects model by brute force by including the dummy variables in the model (assuming it could be done) would induce the biases of the incidental parameters problem, the authors opted for a simpler strategy. They divided the responses (0 to 10) into "dissatisfied" and "satisfied" and recoded the former 0 and the latter 1, producing a binary choice model. The structure, then, is equivalent to

$$\begin{aligned} y_{it}^* &= \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + \varepsilon_{it}, \\ y_{it} &= j \text{ if } \mu_{j-1} \leq y_{it}^* < \mu_j, j = 0,1,\ldots,10, i = 1,\ldots,N, t = 1,\ldots,T_i, \\ z_{it} &= 1 \text{ if } y_{it} > 7. \end{aligned} \tag{9.11}$$

(The average response on the observed y_{it} in the sample was between 7 and 8.) The transformation is equivalent to the 8th of the 10 possible binary choice models in the Das and van Soest (2000) formulation;

$$\text{Prob}(y_{it} > 7 \mid \mathbf{x}_{it}) = \Lambda(\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{\text{it}} - \mu_7).$$

Once again, the constant μ_7 is absorbed in the individual specific constant term, to produce, as before,

$$\text{Prob}[z_{it} = 1 \mid \mathbf{x}_{it}] = \Lambda[\lambda_i + \boldsymbol{\beta}'\mathbf{x}_{it}]. \tag{9.12}$$

The model was then fit using the same Rasch/Andersen/Chamberlain method noted earlier.

Ferrer-i-Carbonell and Frijters (2004) built on this approach in developing an alternative estimator. In their study, the response variable of interest, from the same GSOEP data set, was "General Satisfaction." One of the shortcomings of the fixed effect binary choice model (whether it is estimated conditionally as suggested above) or unconditionally by computing the full set of coefficients including α_i) is that groups that do not change outcomes in the T_i periods fall out of the sample. For the conditional model,

$$\text{Prob}(z_{i1}=1,z_{i2}=1,\ldots,z_{iT}=1 \mid \Sigma_t z_{it} = T) = 1, \tag{9.13}$$

so the contribution of this observation group i to the log likelihood is zero if z_{it} is always equal to 1. (The same occurs if z_{it} equals zero in every period.) For the brute force approach, the likelihood equation for estimation if α_i for a group in which z_{it} is the same in every period is

$$\partial \log L/\partial \alpha_i = \Sigma_t \; f(\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it}) = 0 \text{ if } z_{it} = 1 \text{ in every period,} \tag{9.14}$$
$$\partial \log L/\partial \alpha_i = \Sigma_t -f[-(\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it}] = 0 \text{ if } z_{it} = 0 \text{ in every period.}$$

The first order condition for estimation of α_i cannot be met with a finite α_i if z_{it} is always one or always zero in every period. For ordered choice data, this is likely to be a frequent occurrence, particularly at the two ends of the distribution. The implication is that the samples used for possibly many of the of the binary choice equations in the Das and van Soest (2000) or the Winkelmann and Winkelmann (1998) estimator will lose many observations.

Ferrer-i-Carbonell and Frijters (2004) [and Frijters, Haisken-DeNew and Shields (2004)] modified the Winkelmann and Winkelmann (1998) approach. Initially, the approach is essentially the same, though it begins with a fixed effect *and* individual specific thresholds;

$$y_{it}^* = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},$$
$$y_{it} = j \text{ if } \mu_{i,j-1} \leq y_{it}^* < \mu_{i,j}, j = 0,\ldots,J; i = 1,\ldots,N; t = 1,\ldots,T_i. \tag{9.15}$$

The ordered logit form is assumed. For each individual, i, in the sample, once again,

$$\text{Prob}[z_{it} = 1 \mid \mathbf{x}_{it}] = \Lambda[\lambda_i + \boldsymbol{\beta}'\mathbf{x}_{it}]. \tag{9.16}$$

The difference here is that z_{it} is defined with respect to an individual specific j_i^*, so

$$z_{it} = 1 \text{ if } y_{it} > j_i^* \text{ and } 0 \text{ otherwise.}$$

(In Winkelmann and Winkelmann's method, $j_i^* = 7$ for all i.) The algorithm for choosing j_i^* efficiently for each individual is given in the paper. (The technical Appendix that describes their method can be downloaded from the website for the Royal Economic Society at `http://www.res.org.uk/economic/ta/pdfs/ecoj_235_app.pdf`. It is not contained in the paper, itself.) The resulting contribution to the likelihood for individual i is

$$\text{Prob}(y_{i1} > j_i^*, y_{i2} > j_i^*, \ldots, y_{i,T_i} > j_i^* \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,T_i}, \Sigma_{t=1}^{T_i} z_{it} = c_i)$$
$$= \frac{\exp\left[\Sigma_{t=1}^{T_i} z_{it}\boldsymbol{\beta}'\mathbf{x}_{it}\right]}{\sum_{(z_1,z_2,\ldots,z_{Ti})\in S(j_i^*,c_i)} \exp\left[\Sigma_{t=1}^{T_i} z_t\boldsymbol{\beta}'\mathbf{x}_{it}\right]}. \tag{9.17}$$

Where $c_i = \Sigma_t z_{it}$ = the number of times y_{it} is greater than the chosen threshold. The threshold j_i^* is chosen so that c_i is not equal to 0 or T_i. $S(j_i^*,c_i)$ is the set of all possible vectors, $(z_1,z_2,\ldots,z_{Ti})$, whose elements are all zero or one and sum to c_i; that is, the set of vectors corresponding to sets of outcomes y_{iy} such that c_i of them are greater than j_i^*. The denominator of the probability is the sum over all possible arrangements of T_i z's such that the sum is c_i. [See Krailo and Pike (1984) for the computations involved.]

9.2 Ordered Choice Models with Random Effects

Save for an ambiguity about the mixture of distributions in an ordered logit model, a random effects version of the ordered choice model is a straightforward extension of the binary choice case developed by Butler and Moffitt (1982). An interesting application which appears to replicate, but not connect to Butler and Moffitt is Jansen (1990). Jansen estimates the equivalent of the Butler and Moffitt model with an ordered probit model, using an iterated MLE with

quadrature used between iterations. Following Jansen's lead, Crouchley (1995) also designed the equivalent of the common random effects model, but embeds it in a complementary log-log form that allows, at least for his two period model, a closed form expression for the probabilities after the random effect is integrated out. Characteristically, this strand of the iterature emerged completely apart from the social science counterpart, which had, by then, integrated the random effects, panel data model into a variety of single index specifications such as this one.

Crouchley's formulation of the "random-effects ordered response model" is

$$y_{ij} = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_{ij} + \mathbf{b}_i'\mathbf{z}_{ij} + e_{ij},$$

where $\mathbf{b}_i$ is a vector of individual specific random effects, $\mathbf{x}_{ij}$ is a known design matrix, and e_{ij} is the stochastic disturbance. The model is immediately simplified to a single random effect, $\mathbf{b}_i'\mathbf{z}_{ij} = e_i$, which leaves

$$y_{ij} = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_{ij} + e_i + e_{ij},\ i = 1,...,N,\ j = 1,...T_i. \tag{9.18}$$

The remainder of the treatment is an ordered complementary log-log model with random effects, which is very similar to the model we have considered so far. The difference from this point forward is in the functional form of the distributions of both e_i and e_{ij}, neither of which is assumed to be normal. Crouchley notes, the simplified dimensions can be relaxed

The structure of the random effects ordered choice model is

$$\begin{aligned}
y_{it}^* &= \boldsymbol{\beta}'\mathbf{x}_{it} + u_i + \varepsilon_{it},\\
y_{it} &= \text{j if } \mu_{j-1} \le y_{it}^* < \mu_{it},\\
\varepsilon_{it} &\sim f(.) \text{ with mean zero and constant variance 1 or } \pi^2/3 \text{ (probit or logit)},\\
u_i &\sim g(.) \text{ with mean zero and constant variance, } \sigma^2, \text{ independent of } \varepsilon_{it} \text{ for all } t.
\end{aligned}$$

If we maintain the ordered probit form and assume as well that u_i is normally distributed, then, at least superficially, we can see the implications for the estimator of ignoring the heterogeneity. Using the usual approach,

$$\begin{aligned}
\text{Prob}(y_{it} = j|\mathbf{x}_{it}) &= \text{Prob}(\boldsymbol{\beta}'\mathbf{x}_{it} + u_i + \varepsilon_{it} < \mu_j) - \text{Prob}(\boldsymbol{\beta}'\mathbf{x}_{it} + u_i + \varepsilon_{it} < \mu_{j-1})\\
&= \Phi\left(\frac{\mu_j}{\sqrt{1+\sigma^2}} - \frac{\boldsymbol{\beta}'\mathbf{x}_{it}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\mu_{j-1}}{\sqrt{1+\sigma^2}} - \frac{\boldsymbol{\beta}'\mathbf{x}_{it}}{\sqrt{1+\sigma^2}}\right) \qquad (9.19)\\
&= \Phi(\tau_j - \boldsymbol{\gamma}'\mathbf{x}_{it}) - \Phi(\tau_{j-1} - \boldsymbol{\gamma}'\mathbf{x}_{it}).
\end{aligned}$$

Unconditionally, then, the result is an ordered probit in the scaled threshold values and scaled coefficients. Evidently, this is what is estimated if the data are pooled and the heterogeneity is ignored. (See Wooldridge (2002). Note that a "robust" covariance matrix estimator does not redeem the estimator.)

The likelihood function for a sample can be estimated using the method of Butler and Moffitt. It is convenient to write $u_i = \sigma v_i$ where v_i is the standardized variable – for the moment, N(0,1). Then, conditioned on v_i, the observations on y_{it}, $t = 1,...,T_i$ are independent, so the contribution to the conditional likelihood for individual i would be the joint probability,

$$\text{Prob}(y_{i1} = j_1, y_{i2} = j_2, ..., y_{iT} = j_T|\mathbf{X}_i, v_i) = \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right)\right]. \tag{9.20}$$

The unconditional probability would be, then,

$$P(\mathbf{y}_i = \mathbf{j}_i \mid \mathbf{X}_i) = \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right)\right]\phi(v_i)dv_i, \quad (9.21)$$

(where we have defined a shorthand for the joint probability). The unconditional log likelihood is

$$\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right)\right]\phi(v_i)dv_i. \quad (9.22)$$

The remaining complication is how to compute the integral. Two methods are available. The method of Gauss-Hermite quadrature developed by Butler and Moffitt (1982) uses an approximation to the integrals;

$$\log L_H = \sum_{i=1}^{N} \log \sum_{m=1}^{M} WT_m \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma N_m\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma N_m\right)\right], \quad (9.22)$$

where WT_m and N_m are the weights and nodes, respectively, for the quadrature. [See, e.g., Abramovitz and Stegun (1971).] The accuracy of the approximation is a function of M, the number of quadrature points. Greater accuracy is achieved with increased M, but at the cost of greater computation time. [See, e.g., Rabe-Hesketh, Skrondal and Pickles (2005).] An alternative approach to the estimation would be maximum simulated likelihood. The integral in the log likelihood is

$$\int_{v_i} \left(L_i \mid v_i\right)\phi(v_i)dv_i = E_{v_i}\left[\prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_i\right)\right]\right], \quad (9.23)$$

which can be approximated using simulation. The simulated log likelihood to be maximized is

$$\log L_S = \sum_{i=1}^{N} \log \frac{1}{R}\sum_{r=1}^{R} \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_{ir}\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - \sigma v_{ir}\right)\right], \quad (9.24)$$

where v_{ir}, $r = 1,...,R$ is a set of random draws from the standard normal population (the same set, reused every time the function is calculated for individual i). [See Train (2003) and Greene (2008a, Chapter 17) for details on simulation based estimation.] Neither method of computation has an obvious advantage in this one dimensional integration problem. (In terms of computational time, the advantage shifts significantly in favor of simulation when the number of dimensions (the order of the integration) increases past two.)

The random effects model extends naturally to the ordered probit model if the heterogeneity is viewed as the sum of small influences – a central limit theorem could be invoked to justify the layering of the normally distributed heterogeneity, u_i, on the normally distributed disturbance, ε_{it}. That does raise an ambiguity in the specification of the ordered logit model. The appeal of the logistic distribution is largely its mathematical convenience, though the slightly thicker tails might lend it some additional utility. However, the mixture of a logistic disturbance with a normally distributed random effect is a bit unnatural. The Butler and Moffitt method does not extend readily to integrating the logistic distribution. However, the simulation method can easily be so adapted. The simulated ordered logit model is obtained by using the logistic cdf, $\Lambda(.)$ rather then the normal, $\Phi(.)$ in the function. Draws from the desired distribution are simply obtained by the appropriate transformation of draws, U_{ir}, from the standard uniform, $U(0,1)$; $\Phi^{-1}(U_{ir})$ for simulation from the normal, or $\log[U_{ir}/(1-U_{ir})]$ for the logistic. The optimization process is the same for the two cases. The deeper question would seem to be whether the

logistic/logistic model is a reasonable one in the abstract, compared to the more commonly used normal/normal.

9.3 Testing for Random or Fixed Effects: A Variable Addition Test

A natural question is whether there is a test one can use to determine whether fixed or random effects should be the preferred model. Since the models are not nested, no simple test based on the likelihood function is available. A counterpart to the Hausman (1978) test for the linear model seems desirable, however, unlike the linear case, the fixed effects estimator for this nonlinear model is inconsistent even when it is the appropriate estimator (due to the incidental parameters problem). If one is going to base any test on the estimator of the fixed effects model, it would appear to be necessary to use one of the modified approaches, by Das and van Soest (2000) or Frijters et al. (2004), or any of the individual implied binary choice models, any of which will produce a consistent estimator of β under the hypothesis that the fixed effects model is appropriate. As such, this will force the fixed effects benchmark in the test to rely on the ordered logit model estimates, say $\hat{\beta}_{FE,logit}$. Frijters et al. (2004) argue that the alternative estimator based on a random effects probit specification should estimate a multiple of the same coefficient vector, so the working hypothesis would be $\hat{\beta}_{FE,logit} = \alpha\hat{\beta}_{RE,probit}$. They then propose a type of likelihood ratio test based on computation of the log likelihood functions for the two models. There are a number of problems with this approach, not least of which is that if the working hypothesis is true, it is necessary to estimate α. However, the models are not nested, the parameters must necessarily be based on different sized samples and it is unclear what one should use for the degrees of freedom of the test if it were valid – the authors suggest *K*, the number of parameters in the model, but neither log likelihood forces *K* constraints on the other; the degrees of freedom for the *LR* test is the reduction in the number of dimensions of the parameter space. In this instance, the parameter space has *K* dimensions under both null and alternative. D'Addio, Eriksson and Frijters (2007) estimated a fixed effects ordered logit model and a random effects ordered probit model for "job satisfaction" for data from the European Community Household Panel and found that the fixed effects model was the preferred specification.

No other clearly appropriate procedure has been proposed. This problem is common to other nonlinear models. One strategy does suggest itself, based on the logic of the variable addition test [Wu (1973) and Baltagi (2007)]. In the random effects model to which we added the group means of the variables, the ostensible purpose of the variable addition was to account for correlation between the common effect, u_i, and the regressors. With that correlation present, the appropriate approach is fixed effects. Without that correlation, the random effects model is appropriate. Thus, while conceding that the power of the test is completely unknown at this point, we propose a simple likelihood ratio – variable addition test of the joint significance of the group means in the expanded random effects model.

Estimates of the fixed and random effects models are shown in Tables 9.1-9.3. For our estimated models we have logL = -32656.89 for the random effects model (Table 9.2) and -32588 for the RE model with the group means added (Table 9.3). The likelihood ratio statistic for the hypothesis that the coefficients on the means are all zero is twice the difference, or 137.00, with 5 degrees of freedom. The hypothesis is decisively rejected, so we conclude that the fixed effects model is the preferred specification. Unfortunately, this now raises the question of how to fit the model. The average group size is less than 5. The results in Figure 8.1 suggest that the bias in the full MLE is as much as 30%. The results in Table 9.3 may be the appropriate ones.

Table 9.1 Fixed Effects Ordered Logit Models

Ordered Probability Model		FIXED EFFECTS Ordered Probit Model	
Number of observations	27326	Number of observations	27326
Log likelihood function	-35853.13	Log likelihood function	-28818.86
Number of paameters	9	Number of parameters	5264
Info. Criterion: AIC =	2.62476	Info. Criterion: AIC =	2.49454
Restricted log likelihood	-36734.32	Unbalanced panel has 7293 individuals	
Underlying probabilities based on Logit		2037 groups with inestimable a(i)	

Pooled Estimates					Full Maximum Likelihood Fixed Effects			
Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
Constant	3.6715	.0825	44.526	.0000				
AGE	-.0355	.0011	-30.868	.0000	-.1283	.0057	-22.688	.0000
EDUC	.0625	.0051	12.325	.0000	.0182	.0539	.337	.7360
INCOME	.4592	.0672	6.838	.0000	.4902	.1442	3.400	.0007
MARRIED	.0359	.0298	1.207	.2274	.1085	.0823	1.318	.1876
KIDS	.0971	.0265	3.662	.0003	-.1549	.0577	-2.686	.0072
Mu(1)	2.1671	.0149	145.689	.0000	3.5586	.0490	72.568	.0000
Mu(2)	4.3514	.0151	289.916	.0000	7.1596	.0602	118.857	.0000
Mu(3)	5.1812	.0190	272.983	.0000	8.5189	.0646	131.820	.0000
Conditional Fixed Effects Logit, Healthy=1[Health>2]; Mean=.2288, S.D.=.4200								
AGE	-.1720	.0086	-19.977	.0000				
EDUC	.0213	.0752	.283	.7773				
INCOME	.4931	.2090	2.360	.0183				
MARRIED	.1806	.1147	1.574	.1155				
KIDS	-.0584	.0817	-.714	.4751				

Table 9.2 Random Effects Ordered Logit Models – Quadrature and Simulation

Random Effects Ordered Prob. Model		Random Coefficients OrdProbs Model	
Number of observations	27326	Number of observations	27326
Log likelihood function	-32656.89	Log likelihood function	-32669.96
Info. Criterion: AIC =	2.39090	Info. Criterion: AIC =	2.39186
Unbalanced panel has 7293 individuals		Unbalanced panel has 7293 individuals	

Quadrature based estimation					Max. Sim. Likelihood; 100 Halton draws			
Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
Constant	5.8248	.1690	34.460	.0000	5.7869	.0939	61.617	.0000
AGE	-.0602	.0021	-28.660	.0000	-.0594	.0012	-48.162	.0000
EDUC	.0830	.0113	7.355	.0000	.0838	.0054	15.478	.0000
INCOME	.2664	.0950	2.803	.0051	.2545	.0694	3.673	.0002
MARRIED	.1288	.0473	2.721	.0065	.1225	.0305	4.018	.0001
KIDS	.0147	.0396	.372	.7097	.0158	.0274	.576	.5648
Mu(01)	3.0227	.0358	84.522	.0000	3.0155	.0328	91.954	.0000
Mu(02)	6.2878	.0447	140.610	.0000	6.2824	.0405	154.953	.0000
Mu(03)	7.4514	.0473	157.460	.0000	7.4447	.0430	173.220	.0000
Sigma	1.7935	.0242	74.016	.0000	1.8112	.0153	118.407	.0000

9.4 Extending Parameter Heterogeneity Models to Ordered Choices

Based on the results of the previous sections, the extension of the models with parameter heterogeneity involves only a minor change in the log likelihood and essentially none in the interpretation of the model. For example, in the random parameters model, the heterogeneity in the parameters is the same as in the random effect – it is useful to view the random effects model as a random parameters model in which only the constant term is random. The more general model is

$$\log L_i = \log \int_{\mathbf{w}_i} \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i - (\mathbf{L}\mathbf{D}\mathbf{w}_i)'\mathbf{x}_i\right)\right] f(\mathbf{w}_i) d\mathbf{w}_i. \tag{9.25}$$

The log likelihood for the sample is once again the sum over the N joint observations. The integration can now be replaced with a simulation over R draws from the multivariate standard normal population. The simulated log likelihood is, then

$$\log L = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \left[\Phi\left(\mu_j - \boldsymbol{\beta}'\mathbf{x}_{it} - (\mathbf{L}\mathbf{D}\mathbf{w}_{ir})'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - (\mathbf{L}\mathbf{D}\mathbf{w}_{ir})'\mathbf{x}_i\right)\right]. \tag{9.26}$$

The generalized ordered choice model (3) and the latent class model are handled similarly. For the first,

$$\log L_S(\boldsymbol{\beta},\boldsymbol{\Delta},\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{L},\mathbf{D},\boldsymbol{\sigma},\tau) = \sum_{i=1}^{n} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \left(F\left[\frac{\mu_{ij,r} - \boldsymbol{\beta}_{i,r}'\mathbf{x}_{it}}{\sqrt{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_{i,r})}} \right] - F\left[\frac{\mu_{i,j-1,r} - \boldsymbol{\beta}_{i,r}'\mathbf{x}_{it}}{\sqrt{\exp(\boldsymbol{\gamma}'\mathbf{h}_i + \tau e_{i,r})}} \right] \right). \tag{9.27}$$

As before, the structure assumes that the heterogeneity is constant through time. Bhat (1999) applied a variant of this model to the number of stops on the evening commute in a survey of San Francisco Bay area commuters. The model in this application combined the random effects model (with a heteroscedastic random effect) and the random parameters specification. The model is, thus, the one above with $\boldsymbol{\Delta} = \mathbf{0}$, and $\mathbf{L} = \mathbf{I}$ (the parameters were uncorrelated).

For the latent class model, the appropriate log likelihood function is

$$\log L = \sum_{i=1}^{N} \log \left\{ \sum_{q=1}^{Q} \frac{\exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)}{\sum_{q=1}^{Q} \exp(\theta_q + \boldsymbol{\delta}_q'\mathbf{z}_i)} \left(\prod_{t=1}^{T_i} \sum_{j=0}^{J} m_{ij} \left[\Phi\left(\mu_{j,q} - \boldsymbol{\beta}_q'\mathbf{x}_{it}\right) - \Phi\left(\mu_{j-1,q} - \boldsymbol{\beta}_q'\mathbf{x}_{it}\right) \right] \right) \right\}. \tag{9.28}$$

The counterpart to the assumption of time invariant heterogeneity is the assumption that the class membership is the same in every period.

Random parameters and latent class estimates for the health care model are shown in Tables 9.4-9.6. The latent class model is fit with the full panel data set in Table 9.5, then with the cross section used previously (4,483 observations) in Table 9.6. The estimates are relatively stable across the two samples. However, the benefit from the larger sample is clearly visible in the much smaller standard errors in Table 9.5.

Table 9.3 Random Effects Ordered Probit Model with Mundlak Correction

Random Effects Ordered Probability Model		
Log likelihood function	-32588.39	-32656.89
Info. Criterion: AIC =	2.38625	2.39090

Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
Constant	5.0109	.1890	26.514	.0000	5.8248	.1690	34.460	.0000
AGE	-.1057	.0047	-22.587	.0000	-.0602	.0021	-28.660	.0000
EDUC	.0204	.0548	.373	.7095	.0830	.0113	7.355	.0000
INCOME	.3893	.1212	3.213	.0013	.2664	.0950	2.803	.0051
MARRIED	.0995	.0706	1.408	.1591	.1288	.0473	2.721	.0065
KIDS	-.1249	.0507	-2.465	.0137	.0147	.0396	.372	.7097
AGEBAR	.0591	.0053	11.140	.0000				
EDUCBAR	.0630	.0559	1.127	.2596				
INCBAR	.6255	.2030	3.082	.0021				
MARRBAR	-.1175	.0989	-1.188	.2348				
KIDSBAR	.3559	.0871	4.087	.0000				
Mu(01)	3.0262	.0357	84.767	.0000	3.0227	.0358	84.522	.0000
Mu(02)	6.3001	.0448	140.742	.0000	6.2878	.0447	140.610	.0000
Mu(03)	7.4699	.0475	157.343	.0000	7.4514	.0473	157.460	.0000
Sigma	1.7909	.0241	74.220	.0000	1.7935	.0242	74.016	.0000

Table 9.4 Random Parameters Ordered Logit Model

Ordered LOGIT probability model		
Number of observations	27326	Nonrandom Parameters Model
Log likelihood function	-32895.56	Log likelihood = -35853.13
Info. Criterion: AIC =	2.40874	
Unbalanced panel has	7293 individuals.	
Simulation based on	50 Halton draws	

Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
Means for random parameters					Fixed Parameters			
Constant	5.4942	.0884	62.127	.0000	3.6715	.0825	44.526	.0000
AGE	-.0577	.0012	-48.892	.0000	-.0355	.0011	-30.868	.0000
EDUC	.0980	.0053	18.491	.0000	.0625	.0051	12.325	.0000
INCOME	.2042	.0675	3.027	.0025	.4592	.0672	6.838	.0000
MARRIED	.1582	.0290	5.461	.0000	.0359	.0298	1.207	.2274
KIDS	-.0010	.0267	-.036	.9717	.0971	.0265	3.662	.0003
Std. Devs. of random parameters								
Constant	.0392	.0123	3.197	.0014				
AGE	.0256	.0003	89.092	.0000				
EDUC	.1045	.0012	90.135	.0000				
INCOME	.0425	.0292	1.453	.1463				
MARRIED	.2892	.0132	21.861	.0000				
KIDS	.5574	.0183	30.525	.0000				
Threshold parameters								
MU(1)	2.9688	.0315	94.113	.0000	2.1671	.0149	145.689	.0000
MU(2)	6.1766	.0394	156.656	.0000	4.3514	.0150	289.916	.0000
MU(3)	7.3133	.0420	174.328	.0000	5.1812	.0190	272.983	.0000

Table 9.5 Latent Class Ordered Logit Models

Latent Class / Panel OrdProbs Model					Number of observations 4483			
Number of observations 27326					Log likelihood Func. -5743.560			
Log likelihood function -32639.79					Info. Criterion: AIC = 2.57799			
Info. Criterion: AIC = 2.39148								
Unbalanced panel has 7293 individuals.								
LHS variable = values 0,1,..., 4								
	Panel Data, 7,293 Individuals				Cross Section, 4483 Observations			
Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
Model parameters for latent class 1								
Constant	6.3342	.2374	26.677	.0000	4.8481	1.1057	4.385	.0000
AGE	-.0591	.0024	-24.266	.0000	-.0643	.0151	-4.263	.0000
EDUC	.1067	.0100	10.681	.0000	.0931	.0706	1.319	.1872
INCOME	.2600	.1309	1.987	.0470	.5364	.5847	.918	.3589
MARRIED	.1397	.0545	2.564	.0103	-.8097	1.2774	-.634	.5261
KIDS	-.0024	.0483	-.051	.9595	.1736	.4600	.377	.7059
MU(1)	3.6647	.1566	23.395	.0000	1.6653	.4927	3.380	.0007
MU(2)	7.2343	.1690	42.799	.0000	4.6231	.5832	7.926	.0000
MU(3)	8.6861	.1813	47.901	.0000	6.6873	1.1739	5.697	.0000
Model parameters for latent class 2								
Constant	2.6680	.1756	15.194	.0000	4.8928	1.4105	3.469	.0005
AGE	-.0486	.0026	-18.894	.0000	-.0647	.0242	-2.676	.0074
EDUC	.0688	.0110	6.261	.0000	.0188	.0950	.198	.8427
INCOME	.7206	.1479	4.871	.0000	.8615	.9192	.937	.3486
MARRIED	.2207	.0606	3.640	.0003	1.1373	2.0288	.561	.5751
KIDS	.0234	.0575	.406	.6845	-.4177	.6478	-.645	.5191
MU(1)	2.6404	.0417	63.345	.0000	1.1750	.7604	1.545	.1223
MU(2)	5.0168	.0826	60.747	.0000	4.6922	1.1750	3.993	.0001
MU(3)	5.5344	.1012	54.687	.0000	6.0958	1.3305	4.582	.0000
Model parameters for latent class 3								
Constant	6.1949	.2778	22.299	.0000	4.7151	1.4752	3.196	.0014
AGE	-.0361	.0026	-13.961	.0000	-.0137	.0213	-.640	.5222
EDUC	.0571	.0139	4.119	.0000	-.0564	.0754	-.748	.4542
INCOME	-.4401	.1324	-3.325	.0009	.2615	.6820	.383	.7014
MARRIED	-.0120	.0604	-.198	.8428	.1501	.4114	.365	.7152
KIDS	.0084	.0566	.148	.8821	-.3081	.3045	-1.012	.3117
MU(1)	2.2870	.1839	12.437	.0000	4.0424	1.8974	2.130	.0331
MU(2)	4.6390	.1915	24.229	.0000	4.5358	1.2916	3.512	.0004
MU(3)	5.6626	.1910	29.648	.0000	4.5358	1.2351	3.672	.0002
Estimated prior probabilities for class membership								
ONE_1	-.5926	.1075	-5.514	.0000	.2715	1.7062	.159	.8736
FEMALE_1	-.0311	.0893	-.348	.7276	.1884	.3109	.606	.5446
HANDDU_1	-.7248	.1673	-4.331	.0000	-.3633	.3648	-.996	.3193
WORKIN_1	-.0687	.0960	-.716	.4742	.6391	.3820	1.673	.0944
ONE_2	-.7473	.1026	-7.281	.0000	.5073	1.5932	.318	.7502
FEMALE_2	.2239	.0868	2.579	.0099	-.1693	.3832	-.442	.6588
HANDDU_2	1.1296	.1081	10.447	.0000	-.5219	.3961	-1.317	.1877
WORKIN_2	-.3003	.0906	-3.315	.0009	.1720	.4867	.353	.7238
ONE_3	.00	...(Fixed Parameter)...			.00	...(Fixed Parameter)..		
FEMALE_3	.00	...(Fixed Parameter)...			.00	...(Fixed Parameter)..		
HANDDU_3	.00	...(Fixed Parameter)...			.00	...(Fixed Parameter)..		
WORKIN_3	.00	...(Fixed Parameter)...			.00	...(Fixed Parameter)..		
Prior class probabilities at data means for LCM variables								
	Class 1	Class 2	Class 3		Class 1	Class 2	Class 3	
	.50172	.22339	.27489		.44728	.34178	.21094	

9.5 Dynamic Models

Dynamic effects in ordered choice models have been introduced in two settings. In the pure time series applications in which researchers have examined asset price movements, interest rate changes and monetary policy, the focus is on inertia, and takes the form of an autoregressive model in the latent variable regression. The Czado, Heyn and Müller (2005) and Müller and Czado (2005) study of migraine headache severity is also presented in this framework, though their study can be usefully viewed as falling somewhere between the time series analysis of, e.g., Eichengreen et al.'s (1985) study of bank rate policy and the recent panel data studies, e.g., of health satisfaction. In panel data settings, such as Contoyannis, Jones and Rice (2004), the model is directed at state dependence, and, instead, takes the form of lagged effects in the observed variables. We will examine each of these in a bit more detail.

A natural form of the ordered probit model with lagged effects is suggested by Girard and Parent (2001),

$$
\begin{aligned}
y_t^* &= \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t, \\
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t, \\
y_t &= j \text{ if } \mu_{j-1} < y_t^* \leq \mu_j,
\end{aligned}
$$

with the usual restrictions. Estimation is carried using a Gibbs sampler (MCMC) and using Albert and Chib's data augmentation method; the values y_t^* as well as the initial value, y_0^* are treated as nuisance parameters to be included with $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and ρ for posterior analysis.

Eichengreen, Watson and Grossman (1985) examined the Bank Rate (BR) adjustment policies of the Bank of England over a period of 328 weeks. The structural model is

$$
\begin{aligned}
\text{Prob}[\Delta BR_t = -50 \mid J_t] &= P_{1t}(J_t), \\
\text{Prob}[\Delta BR_t = 0 \mid J_t] &= P_{2t}(J_t), \\
\text{Prob}[\Delta BR_t = 100 \mid J_t] &= P_{3t}(J_t), \; t = 1,\ldots,T,
\end{aligned}
$$

where the adjustment rates are in basis points and J_t is an information set that contains current and lagged values of exogenous variables $\mathbf{x}_t$ and the entire preceding history of bank rates, BR_s, $s = 1,\ldots,t\text{-}1$. An underlying regression is specified for the "change in an unobserved "underlying" bank rate,

$$
\Delta BR_t^* = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t, \; \varepsilon_t \mid J_t \sim \text{N}[0,\sigma^2].
$$

The observed Bank Rate changes when it is too far from BR_t^* according to the rule,

$$
\begin{aligned}
\Delta BR_t &= -50 \text{ if } BR_t^* < BR_{t-1} - \alpha_L, \\
\Delta BR_t &= 0 \text{ if } BR_{t-1} - \alpha_L < BR_t^* < BR_{t-1} + \alpha_U, \\
\Delta BR_t &= 100 \text{ if } BR_t^* > BR_{t-1} + \alpha_U.
\end{aligned}
$$

Thus, the rule is that the observed rate decreases by 50 basis points if BR_t^* is "appreciably" less than BR_{t-1} and increases by 100 basis points if BR_t^* is appreciably greater than BR_{t-1}. Appreciably is defined by the unknown threshold values, α_L and α_U. The authors note, the model resembles a familiar ordered probit model, but differs in at least two major respects. First, although the structural equations describe the changes in *BR*, the inequalities that invoke the similarity with the ordered probit model are defined in the levels of *BR*, not changes. Thus, there are stochastic dynamics in BR_t. Second, since the lagged value of the observed time series appear in the model definition, the identification of the model parameters must be developed in detail. It does not

follow from simple examination of the specification as it does in the conventional model. The likelihood function (see their pp. 741-744) is markedly more complicated than that we have examined so far. Among the most challenging aspects is that because of the autoregressive nature of the random components in the model, the time series must be treated as a single *T*-variate observation. That implies integration of a *T* (=328) variate normal integral. A strategy is devised in the paper. Eichengreen et al.'s (1985) study has provided the foundation for a number of subsequent studies of bank policy, including Genberg and Gerlach (2004) and Basu and de Jong (2006).

A somewhat simpler form of the ordered probit model has been used to analyze movements in stock prices when the movements of an underlying continuous price variable are expressed in discrete units ("ticks"). Tsay (2002) presents the following general characterization of an application: Define y_{it}^* to be the unobservable true price change of an asset, so that

$$y_{it}^* = P_{it}^* - P_{i,t-1}^*,$$

where P_{it}^* is the virtual price of the asset at time t. The ordered probit model derives from the assumed structure

$$\begin{aligned} y_{it}^* &= \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}, \\ \mathrm{E}[\varepsilon_{it}|\mathbf{x}_{it},\mathbf{w}_{it}] &= 0, \\ \mathrm{Var}[\varepsilon_{it}|\mathbf{x}_{it},\mathbf{w}_{it}] &= \sigma^2(\mathbf{w}_{it}), \\ \varepsilon_{it}|\mathbf{x}_{it},\mathbf{w}_{it} &\sim \mathrm{N}[0,\sigma^2(\mathbf{w}_{it})], \end{aligned}$$

where $\mathbf{x}_{it}$ might contain the exogenously determined information available at time t-1 and $\mathbf{w}_{it}$ is conditioning data such as the time interval of the change as well as "some conditional[ly] heteroscedastic variables." If the observed price change is restricted to a fixed set of intervals, then an ordered probit model emerges;

$$y_{it} = s_j \text{ if } \alpha_{j-1} < y_{it}^* \leq \alpha_j, j = 1,\ldots,J.$$

What follows is a familiar ordered probit model, distinguished from our earlier model by the assumed heteroscedasticity of ε_{it}. Tsay describes in detail an early study of more than 100 stocks by Hausman, Lo and MacKinlay (1992). Hausman et al. describe three features of the American stock market that motivate their treatment: First, stock prices were stated at the time (no longer) in discrete, 1/8 dollar units, so the true continuous variable could not be measured. Second, the timing of transactions can be irregular and random, which makes discrete time modeling problematic. Third, received models have not adequately accounted for the correlations between price changes and other economic variables – these are captured in the latent regression equation in the ordered probit model.

Czado, Heyn and Müller (2005) also used a time series model with dynamics in the latent variable to study the reported severity of migraine headaches reported in the diary of a single patient. The underlying variable, severity of the headache in interval t, is modeled as

$$y_t^* = \boldsymbol{\beta}'\mathbf{x}_t + \gamma y_{t-1}^* + \varepsilon_t.$$

The observed severity is recorded on a scale 0,1,…,5, four times per day over a period of 268 days. The regressor variable includes such variables as weather conditions and day of the week. The application is a pure time series model. As in the Eighengreen et al. study, the dynamics greatly complicate the estimation process. A customized form of Markov Chain Monte Carlo (Bayesian) estimation method for this model is presented in Müller and Czado (2005).

The autoregressive models examined so far are natural specifications for the observed outcomes. Contoyannis, Jones and Rice (2004) examined self assessed health status in the British Household Panel Survey (BHPS). The measure of health status is reported with values 1,…,5. Individuals have a general tendency to repeat the same value unless other factors change. The common effects regression suggested to account for this state dependence is

$$h_{it}^* = \boldsymbol{\beta}'\mathbf{x}_{it} + \Sigma_{j=1}^{5} \gamma_j m_{j,i,t-1} + \alpha_i + \varepsilon_{it},$$

where α_i is a fixed effect and

$$m_{i,j,t-1} = 1 \text{ if } y_{i,t-1} = j \text{ and 0 otherwise.}$$

A familiar ordered probit model applies;

$$h_{it} = j \text{ iff } \mu_{j-1} < y_{it}^* \leq \mu_j.$$

Initially, it is proposed to treat this as a random effects model using the method of Butler and Moffitt (1982). In order to accommodate possible correlation between α_i and the (means of the) other variables and to handle the problem of the initial conditions [Heckman (1981)], they employ the Mundlak (1978) device in:

$$\alpha_i = \alpha_0 + \Sigma_{j=1}^{5} \alpha_j m_{i,1,j} + \boldsymbol{\theta}'\bar{\mathbf{x}}_i + u_i.$$

where $u_i \sim N[0,\sigma^2]$. Inserting this equation into the latent regression provides their ordered probit model,

$$h_{it}^* = \boldsymbol{\beta}'\mathbf{x}_{it} + \Sigma_{j=1}^{5} \gamma_j m_{j,i,t-1} + \alpha_0 + \Sigma_{j=1}^{5} \alpha_j m_{i,1,j} + \boldsymbol{\theta}'\bar{\mathbf{x}}_i + u_i + \varepsilon_{it}.$$

(A few normalizations, such as removal of a redundant constant term, are needed to secure identification of the parameters.) A final adjustment to the model based on a procedure devised by Wooldridge (2002a) is used to account for the rather substantial attrition over the 8 waves of their panel.

10

Bivariate and Multivariate Ordered Choice Models

The preceding sections have examined the more or less standard approaches to modeling ordered data, beginning with the most basic model and ending with various specifications that accommodate observed and unobserved heterogeneity in panel data. In what follows, we examine some recent extensions of the model that include modifications to the basic structure and additions to it that occasionally mandate multiple equation frameworks. It will emerge shortly that most of these extensions do not fit comfortably into the ordered logit framework. At this point, it will prove convenient to drop the distinction between the probit and logit models, and focus attention, as in the received literature, on the ordered probit model.

10.1 Multiple Equations

A multiple equation specification for, say, M ordered choices is a natural extension of the model. The extension is based on a seemingly unrelated regressions (SUR) model for the latent regressions:

$$
\begin{aligned}
&y_{i,1}^* = \boldsymbol{\beta}_1'\mathbf{x}_{i,1} + \varepsilon_{i,1},\ y_{i,1} = j \text{ if } \mu_{j-1,1} < y_{i,1}^* < \mu_{j,1},\ \varepsilon_{i,1} \sim N[0,1],\\
&\cdots\\
&y_{i,M}^* = \boldsymbol{\beta}_M'\mathbf{x}_{i,M} + \varepsilon_{i,M},\ y_{i,M} = j \text{ if } \mu_{j-1,M} < y_{i,M}^* < \mu_{j,M},\ \varepsilon_{i,M} \sim N[0,1],\\
&(\varepsilon_{i,1},\ldots,\varepsilon_{i,M}) \sim N[\mathbf{0},\mathbf{R}],
\end{aligned}
$$

where $\mathbf{R}$ is the unrestricted correlation matrix of the random terms. In principle, this is a straightforward extension of the single variable model. The estimation is substantially complicated because of the amount of computation involved. In the one variable case, the probability is the area under the univariate normal density bounded by two points on a line, which requires two function evaluations of the univariate normal cdf. For two dimensions, the probability is the volume under the bivariate normal surface bounded by a rectangle, which, in general, requires four function evaluations of the bivariate normal integral. For three dimensions, it requires eight function evaluations of the trivariate normal integral. And so on. The amount of computation rises with 2^M. Moreover, the computation of the integrals, themselves, is cumbersome. For one dimension, the typical library routine computation of the normal integral involves evaluation of a ratio of two fourth or fifth order polynomials. The bivariate normal integral must typically be done using quadrature. [See, e.g., Drezner (1978).] For three dimensions or higher, the computation is generally be done by simulation, which will (with current techmology) involve a formidable amount of computing. [But, see Drezner (1994).] This model, even for only two dimensions, does not lend itself conveniently to the ordered logit form, and the received applications use the ordered probit model exclusively. [See, however, Dardanomi and Forcina (2004), who do obtain some analytical results for a multivariate ordered logit model.]

10.2 Bivariate Ordered Probit Models

The two equation case has dominated the received applications, largely because of the practical difficulty of evaluating the higher order normal integrals needed to estimate the models. For two outcomes, we have

$$\begin{aligned} y_{i,1}^* &= \boldsymbol{\beta}_1'\mathbf{x}_{i,1} + \varepsilon_{i,1},\ y_{i,1} = j \text{ if } \mu_{j-1} < y_{i,1}^* < \mu_j,\ j = 0,\ldots,J_1, \\ y_{i,2}^* &= \boldsymbol{\beta}_2'\mathbf{x}_{i,2} + \varepsilon_{i,2},\ y_{i,2} = j \text{ if } \delta_{j-1} < y_{i,2}^* < \delta_j,\ j = 0,\ldots,J_2, \\ \begin{pmatrix}\varepsilon_{i,1}\\ \varepsilon_{i,2}\end{pmatrix} &\sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]. \end{aligned} \tag{10.1}$$

The joint probability for $y_{i,1} = j$ and $y_{i,2} = k$ is

$$\begin{aligned} &\text{Prob}(y_{i,1} = j, y_{i,2} = k \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}) = \\ &\begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}),(\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}),\rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}),(\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}),\rho] \end{bmatrix} - \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}),(\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}),\rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}),(\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}),\rho] \end{bmatrix}. \end{aligned} \tag{10.2}$$

These are the probabilities that enter the log likelihood for a maximum likelihood estimator of the parameters.

Partial effects for this model will be complicated functions of the parameters regardless of how they are defined. But, for a bivariate model, such as this one, even what margin is of interest is not obvious. Derivatives of the bivariate probability, $\text{Prob}(y_{i,1} = j, y_{i,2} = k \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2})$ might well not correspond to a useful experiment. One might, instead, wish to compute the derivatives of the conditional probability,

$$\text{Prob}(y_{i,1} = j \mid y_{i,2} = k, \mathbf{x}_{i,1}, \mathbf{x}_{i,2}) = \frac{\text{Prob}(y_{i,1} = j, y_{i,2} = k \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2})}{\text{Prob}(y_{i,2} = k \mid \mathbf{x}_{i,2})}. \tag{10.3}$$

The denominator would be computed using the marginal, univariate ordered probit model. In either case, the computation will be based on a common result. For convenience, we drop the observation subscript and define the variables,

$$A_L = \mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_1,\ A_U = \mu_j - \boldsymbol{\beta}_1'\mathbf{x}_1,\ B_L = \delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_2,\ B_U = \delta_k - \boldsymbol{\beta}_2'\mathbf{x}_2,$$

where subscripts "*L*" and "*U*" refer to "lower" and "upper," respectively. Then, the bivariate probability is

$$\text{Prob}(y_{i,1} = j, y_{i,2} = k \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}) = \begin{Bmatrix} [\Phi_2[A_U, B_U, \rho] - \Phi_2[A_L, B_U, \rho]] - \\ [\Phi_2[A_U, B_L, \rho] - \Phi_2[A_L, B_L, \rho]] \end{Bmatrix}, \tag{10.4}$$

and the marginal univariate probability is

$$\text{Prob}(y_2 = k) = \Phi(B_U) - \Phi(B_L). \tag{10.5}$$

Computing partial effects from either viewpoint will require the result

$$\frac{\partial \Phi_2(A,B,\rho)}{\partial A} = \phi(A)\Phi\left(\frac{B-\rho A}{\sqrt{1-\rho^2}}\right). \tag{10.6}$$

(The result is symmetric in A and B.) Collecting results, then

$$\begin{aligned}
\frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1} &= \left(\begin{array}{l} \phi(A_U)\Phi\left(\frac{B_U-\rho A_U}{\sqrt{1-\rho^2}}\right) - \phi(A_L)\Phi\left(\frac{B_U-\rho A_L}{\sqrt{1-\rho^2}}\right) \\ -\phi(A_U)\Phi\left(\frac{B_L-\rho A_U}{\sqrt{1-\rho^2}}\right) + \phi(A_L)\Phi\left(\frac{B_L-\rho A_L}{\sqrt{1-\rho^2}}\right) \end{array} \right)(-\boldsymbol{\beta}_1), \\
\frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_2} &= \left(\begin{array}{l} \phi(B_U)\Phi\left(\frac{A_U-\rho B_U}{\sqrt{1-\rho^2}}\right) - \phi(B_L)\Phi\left(\frac{A_U-\rho B_L}{\sqrt{1-\rho^2}}\right) \\ -\phi(B_L)\Phi\left(\frac{A_U-\rho B_L}{\sqrt{1-\rho^2}}\right) + \phi(B_L)\Phi\left(\frac{A_L-\rho B_L}{\sqrt{1-\rho^2}}\right) \end{array} \right)(-\boldsymbol{\beta}_2).
\end{aligned} \tag{10.7}$$

If any variables appear in both equations, the effects are added. For the conditional probabilities,

$$\begin{aligned}
\frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)/\text{Prob}(y_2 = k \mid \mathbf{x}_2)}{\partial \mathbf{x}_1} &= \frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)/\partial \mathbf{x}_1}{\text{Prob}(y_2 = k \mid \mathbf{x}_2)}, \\
\frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)/\text{Prob}(y_2 = k \mid \mathbf{x}_2)}{\partial \mathbf{x}_2} &= \frac{\partial \text{Prob}(y_1 = j, y_2 = k \mid \mathbf{x}_1, \mathbf{x}_2)/\partial \mathbf{x}_2}{\text{Prob}(y_2 = k \mid \mathbf{x}_2)} \\
&\quad - \text{Prob}(y_1 = j \mid y_2 = k, \mathbf{x}_1, \mathbf{x}_2)\frac{\phi(B_U)-\phi(B_L)}{\text{Prob}(y_2 = k \mid \mathbf{x}_2)}(-\boldsymbol{\beta}_2).
\end{aligned} \tag{10.8}$$

As before, if variables appear in both equations, the two components are added. Before examining the applications of the model in detail, it is useful to look more closely at some special cases.

An admittedly trivial extension is the bivariate model in which ρ equals zero. In this instance, the bivariate model becomes a pair of univariate models. We mention this case at this point, as chronologically, the second application of the bivariate ordered probit model, Gustaffson and Stafford (1992), used this model to study child care subsidies and labor supply behavior for a sample of Swedish mothers. The hypothesis of uncorrelated equations is easily testable in this setting using either a likelihood ratio test or the Wald statistic (t ratio) associated with the estimate of ρ. Butler and Chatterjee (1995) consider other tests of the model specification, normality and exogeneity of the right hand sides, using GMM rather than maximum likelihood estimation. (They apply their methods to the study of dogs/television ownership noted below.) Guo, Bhat and Copperman (2003) used the unrestricted model shown above to model the joint count of motorized and nonmotorized trips for a survey of individuals in the San Francisco Bay area.

10.3 Polychoric Correlation

The *polychoric correlation coefficient* is computed for a pair of discrete ordered variables, such as $y_{i,1}$ and $y_{i,2}$ above. The theory behind the computation is that $y_{i,1}$ and $y_{i,2}$ are censored versions of underlying, bivariate normally distributed variables, again, precisely as $y_{i,1}$ and $y_{i,2}$ above are obtained. The polychoric correlation coefficient is an estimator of the correlation coefficient in the underlying bivariate normal distribution. The best known method of computing the coefficient for grouped data (in the form of contingency tables), is due to Olssen (1979, 1980). [See , also, Ronning (1990) and Ronning and Kukuk (1996).] The development above suggests a counterpart for how to compute the coefficient when the data are individually measured. If the two equations in the bivariate model have only their constant terms, and no regressors, then precisely the suggested underlying model emerges.

$$
\begin{aligned}
&y_{i,1}^* = \beta_1 + \varepsilon_{i,1},\ y_{i,1} = j \text{ if } \mu_{j-1} < y_{i,1}^* < \mu_j, j = 0,\ldots,J_1,\\
&y_{i,2}^* = \beta_2 + \varepsilon_{i,2},\ y_{i,2} = j \text{ if } \delta_{j-1} < y_{i,2}^* < \delta_j,\ j = 0,\ldots,J_2,\\
&\begin{pmatrix}\varepsilon_{i,1}\\ \varepsilon_{i,2}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].
\end{aligned}
\tag{10.9}
$$

Thus, the implied algorithm, which has been built into modern software such as *NLOGIT*, *Stata* and *SAS*, is simply to fit a bivariate ordered probit model which has only constant terms in the two equations. [See, as well, Calhoun (1986, 1995) for further discussion of computer programs.] Returning to the regression model, it follows that the correlation coefficient in the bivariate ordered probit regression model can be interpreted as the conditional (on $\mathbf{x}_{i,1}$ and $\mathbf{x}_{i,2}$) polychoric correlation coefficient.

10.4 Semi-Ordered Bivariate Probit Model

A second interesting special case arises if one of the variables is binary;

$$
\begin{aligned}
&y_{i,1}^* = \boldsymbol{\beta}_1'\mathbf{x}_{i,1} + \varepsilon_{i,1}, y_{i,1} = 0 \text{ if } y_{i,1}^* < 0, \text{ and } y_{i,1} = 1 \text{ if } y_{i,1}^* > 0,\\
&y_{i,2}^* = \boldsymbol{\beta}_2'\mathbf{x}_{i,2} + \varepsilon_{i,2},\ y_{i,2} = j \text{ if } \delta_{j-1} < y_{i,2}^* < \delta_j,\ j = 0,\ldots,J_2,\\
&\begin{pmatrix}\varepsilon_{i,1}\\ \varepsilon_{i,2}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].
\end{aligned}
\tag{10.10}
$$

This case (like the previous one) does not mandate any special modification of the likelihood function. The appropriate terms can be obtained directly from the earlier general result. This particular form has appeared in a number of applications, under the name "Bivariate Semi-Ordered Probit Model." Weiss (1993) used this model to examine the extent of injuries in motorcycle accidents, with the binary variable being helmet use. Armstrong and McVicar (2000) used this form to examine the relationship between education and vocational training for a sample of Irish youth. McVicar and McKee (2002), using the same model, studied the two variables, vocational attainment (ordered) and working part time during education (binary), also for a sample of Irish youth. In this study, the education achievement is a four level exam measure.

10.5 Applications of the Bivariate Ordered Probit Model

The first application of the bivariate ordered probit model is Calhoun (1991, 1994) who examined the joint distribution of "Desired Family Size" (DFS) and "Children Ever Born" (CEB). In a followup analysis, he used CEB to truncate DFS, to eliminate unwanted children, then reexamined the model with this form of truncation. In an application to descriptions of criminal behavior and subsequent labor market experience, Nagin and Waldfogel (1995) examined the job market performance of young British offenders at ages 17 and 19. In a related analysis, Paternoster and Brame (1998) examined "self control" and "criminal behavior" in a study in criminology [See, also, comments in Britt (2000).] Butler and Chatterjee (1997), in their contribution to pet econometrics, analyzed the joint ownership of dogs and televisions. This is one of several studies in which authors used the bivariate ordered probit to model variables that arguably should be analyzed as counts (with something like a Poisson regression model. However, the bivariate Poisson regession model remains to be well developed. [See, also, Sanko et al. (2004) who looked at ownership of cars and motorcycles and Bhat et al. (1996b, 1998, 1999, 2000, 2002) who analyzed vehicle ownership, trip counts and activity counts.] The ordered probit model has been modified for use in contingent valuation studies, in which survey respondents express their preferences with a range of values rather than a point. Kuriama et al. (1998) used a contingent valuation study to examine consumers' preferences for a world heritage site in Japan. The ordered probit study follows a Vote/No Vote choice, and so has elements of the semiordered bivariate probit model described earlier as well. In two very natural application, Kohler and Rodgers (1999) studied the motivation to have children in a survey of pairs of twins. Christensen et al. (2003) also examined twins, in their case, seeking a genetic effect on fertility. Biswas and Das (2002) examined an epidemiologic study of diabetic retinopathy. Separate equations are specified for the right and left eye severity of the disease (coded 0 to 4). This is one of only a few Bayesian applications. [Biswas and Das benchmarked their study against an earlier analysis of the same data by Kim (1995). It is surprising that they did not use Kim's estimates in their priors. This seems like a natural application of Bayesian updating.] A variety of other applications have appeared, most since 2000, in economics, finance and transportation research. Table 9.1 lists some of the recent applications. (Full citations appear in the references list.)

Table 10.1 Applications of Bivariate Ordered Probit Since 2000

Year	Authors	Application
2000	Magee, et al.	Correlation between husband's and wife's education
2000	Bhat and Singh	Bivariate count model travel related activities
2002	Lawrence and Palmer	Views on health care reform,
2003	Guo, Bhat, Copperman	Counts of motorized and nonmotorized trips
2004	Bedi and Tunali	Participation in land and labor contracts in turkish agriculture
2004	Dupor et al.	Federal Reserve Open Market Committee: Bias announcement (ease, neutral, tighten) and magnitude of next meeting adjustment (-25, 25/0, 0, 0/25, 25+)
2005	Dueker et al.	Job restrictions of nurses:
2005	Filer and Honig	Pensions and retirement behavior,
2006	Adams	University and internal cost allocations of R&D expenditure
2006	Scott and Axhausen	Interactions between cars and season tickets,
2006	Scotti	Bivariate Model of Fed and European Central Bank main policy rates
2007	Mitchell and Weale	Accuracy of expectations about financial circumstances in the British Household Panel Survey

10.6 A Panel Data Version of the Bivariate Ordered Probit Model

Since it is a two equation model, it is unclear how common heterogeneity effects should enter the bivariate model. [See, e.g., Verbeek (1990), Verbeek and Nijman (1992) and Zabel (1992) for a similar exchange in the context of the sample selection model.] Generically, a bivariate model with time invariant random effects might appear

$$
\begin{aligned}
y_{it,1}^* &= \boldsymbol{\beta}_1'\mathbf{x}_{it,1} + \varepsilon_{it,1} + u_{1,i}; \quad y_{it,1} = j \text{ if } \mu_{j-1} < y_{it,1}^* < \mu_j, j = 0,\ldots,J_1, \\
y_{it,2}^* &= \boldsymbol{\beta}_2'\mathbf{x}_{it,2} + \varepsilon_{it,2} + u_{2,i}; \quad y_{it,2} = j \text{ if } \delta_{j-1} < y_{it,2}^* < \delta_j, j = 0,\ldots,J_2, \qquad (10.11)
\end{aligned}
$$

$$
\begin{pmatrix} \varepsilon_{it,1} \\ \varepsilon_{it,2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]; \quad \begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right].
$$

Computation of the parameters in this model would involve integration over both bivariate normal integrals. The approach used by Riphahn, Wambach and Million (2003) for a bivariate Poisson model with two random effects suggests a strategy. Conditioned on the random effects, the likelihood function is

$$
L \mid \mathbf{u}_1, \mathbf{u}_2 = \prod_{i=1}^{N} \prod_{t=1}^{T_i} \sum_{j=0}^{J_1} \sum_{k=0}^{J_2} m_{it,j} n_{it,k} \left\{ \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - u_{i1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - u_{i2}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - u_{i1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - u_{i2}), \rho] \end{bmatrix} - \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - u_{i1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - u_{i2}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - u_{i1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - u_{i2}), \rho] \end{bmatrix} \right\}, \qquad (10.12)
$$

where $m_{it,j} = 1$ if $y_{it,1} = j$ and 0 otherwise and $n_{it,k} = 1$ if $y_{it,2} = k$ and 0 otherwise. To obtain a form of the likelihood function we can use for estimation, it is necessary to eliminate the unobserved random effects. We use a Cholesky decomposition of the covariance matrix to write

$$
\begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} = \begin{bmatrix} \gamma_{11} & 0 \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{pmatrix} v_{i,1} \\ v_{i,2} \end{pmatrix}, \qquad (10.13)
$$

where $(v_{i,1}, v_{i,2})$ are independent N(0,1) variables. It follows that $\gamma_{11}^2 = \sigma_1^2$, $\gamma_{21}^2 + \gamma_{22}^2 = \sigma_2^2$ and $\gamma_{21}\gamma_{22} = \sigma_{12}$. The specific probabilities with this substitution become

$$
\begin{aligned}
&\text{Prob}(y_{it,1} = j, y_{it,2} = k \mid u_{i1}, u_{i2}, \mathbf{x}_{it,1}, \mathbf{x}_{it,2}) = \\
&\left\{ \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1} - \gamma_{22}v_{i2}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1} - \gamma_{22}v_{i2}), \rho] \end{bmatrix} - \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1} - \gamma_{22}v_{i2}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1} - \gamma_{22}v_{i2}), \rho] \end{bmatrix} \right\}. \qquad (10.14)
\end{aligned}
$$

The unconditional log likelihood is obtained by integrating out the random effects. This step has been simplified by the Cholesky decomposition, since the bivariate integration involves independent standard normals. This could be done using nested Hermite quadratures or simulation. The latter is likely to be simpler and faster. The simulated log likelihood function is

$$\log L_S = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \sum_{j=0}^{J_1} \sum_{k=0}^{K} m_{it,j} n_{it,k} \left\{ \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1,r}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1,r} - \gamma_{22}v_{i2,r}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1,r}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1,r} - \gamma_{22}v_{i2,r}), \rho] \end{bmatrix} - \begin{bmatrix} \Phi_2[(\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1,r}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1,r} - \gamma_{22}v_{i2,r}), \rho] \\ -\Phi_2[(\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{it,1} - \gamma_{11}v_{i1,r}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{it,2} - \gamma_{21}v_{i1,r} - \gamma_{22}v_{i2,r}), \rho] \end{bmatrix} \right\}. \quad (10.15)$$

A fixed effects model might be considered as an alternative, however this would have several drawbacks: Two full sets of effects must be estimated. As usual, the fixed effects preclude time invariant variables in either equation. Though it remains to be established, it seems likely that the force of the incidental parameters problem (small T bias) would operate here as well. The Mundlak (1978) device of including the group means of the time varying variables in the equations might be a useful middle ground.

10.7 Trivariate and Multivariate Ordered Probit Models

As noted earlier, for practical reasons, the bivariate probit is more or less the dimensional limit of the applications of the multivariate ordered probit model. Nonetheless, there have been a handful of applications of the trivariate probit model. Two in the area of transportation research that focus on joint determination of activity and travel model are Scott and Kanaroglou (2001) and Buliung (2005). Genius, Pantzios and Tzouvelakis (2005) estimate a "trivariate semi-ordered probit model." In their application to organic farming in Greece, two of the three equations, contact with an extension agent and use of other sources of information, are binary, while the land adoption decision (none, part, full) has three outcomes. Crouchley (2005) is a methodology study.

Bhat and Srinivasan (2005) suggested how one might extend the ordered probit models to an arbitrary number of equations. Using our own notation, define the latent seemingly unrelated regressions system with ordered choice equations,

$$\begin{aligned} &y_{i1}^* = \boldsymbol{\beta}_1'\mathbf{x}_{i1} + \varepsilon_{i1} + u_{i1},\ y_{i1} = j \text{ if and only if } \mu_{j\text{-}1,1} < y_{i1}^* \le \mu_{j,1},\ j=0,...,J_1, \\ &y_{i2}^* = \boldsymbol{\beta}_2'\mathbf{x}_{i2} + \varepsilon_{i2} + u_{i2},\ y_{i2} = j \text{ if and only if } \mu_{j\text{-}1,2} < y_{i2}^* \le \mu_{j,2},\ j=0,...,J_2, \\ &\qquad\qquad\vdots \\ &y_{iM}^* = \boldsymbol{\beta}_M'\mathbf{x}_{iM} + \varepsilon_{iM} + u_{iM},\ y_{iM} = j \text{ if and only if } \mu_{j\text{-}1,M} < y_{iM}^* \le \mu_{j,M},\ j=0,...,J_M. \end{aligned} \quad (10.16)$$

The usual restrictions on the threshold parameters are applied for each equation. The random components $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\varepsilon_{i2},...,\varepsilon_{iM})'$ are assumed to be independent, identically distributed with standard logistic distributions. Thus, conditioned on u_{im}, each equation defines an ordered logit model. (The choice of a mixture of logit and probit models in this instance seems difficult to motivate. However, it will be a trivial modification of the model to assume that ε_{im} is normally distributed. This would appear to be a more natural specification.) The additional set of random terms, $\mathbf{u}_i = (u_{i1},u_{i2},...,u_{iM})$ are assumed to be distributed as multivariate normal with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{R}$. (As always, unrestricted variance parameters are unidentified.) Nonzero elements in the correlation matrix reflect the presence of common unobserved factors. Conditioned on $\mathbf{u}_i$, the M observed random variables are independent, so the conditional joint likelihood function for individual i is

$$L_i \mid \mathbf{u}_i = \prod_{m=1}^{M} P(y_{im} \mid \mathbf{x}_{im}, u_{im}) . \tag{10.17}$$

The unconditional likelihood function is obtained by integrating $\mathbf{u}_i$ out of the function,

$$L_i = \int_{\mathbf{u}_i} \prod_{m=1}^{M} P(y_{im} \mid \mathbf{x}_{im}, u_{im}) f_M(\mathbf{u}_i : \mathbf{R}) d\mathbf{u}_i , \tag{10.18}$$

where the integral is over the M dimensional random vector $\mathbf{u}_i$ and $f_M(\mathbf{u}_i : \mathbf{R})$ is the M variate normal density with zero means and correlation matrix $\mathbf{R}$. The log likelihood function would then be obtained by summing the logs of the contributions to the likelihood function. The integrals are not directly computable, so Bhat and Srinivasan (2005) propose to use simulation instead. In order to obtain the estimating equations, the device used in Section 5.2.5 is used to install the unknown correlation matrix directly into the model. We write $\mathbf{u}_i = \mathbf{D}\mathbf{v}_i$ where $\mathbf{D}$ is the lower triangular Cholesky factorization of $\mathbf{R}$, so $\mathbf{R} = \mathbf{D}\mathbf{D}'$ and $\mathbf{v}_i$ has a standard normal distribution, with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. ($\mathbf{D}$ is lower triangular with ones on the diagonal.) Then, the simulated log likelihood function is

$$\log L_D = \sum_{i=1}^{N} \log \frac{1}{Q} \sum_{q=1}^{Q} \prod_{m=1}^{M} P(y_{im} : \boldsymbol{\beta}_m, \boldsymbol{\mu}_m \mid \mathbf{x}_{im}, u_{im,q}) ,$$

Where (10.19)

$$u_{im,q} = \sum_{t=1}^{m-1} D_{i,t} v_{it,q} + v_{im,q} .$$

The simulation is run over Q random (or pseudo-random) draws on $\mathbf{v}_i$. The authors used Halton sequences to accelerate the simulation process. The model was applied to a setting of counts of "stops" in seven activity categories. We note, the interpretation of the coefficients in the model will be problematic. Partial effects could be computed from the equations individually by contitioning on then integrating out the random factors as done above for the likelihood function. However, there is no comparability of the coefficients across equations, as each equation has its own scale factor,

$$\text{Var}[\varepsilon_{im} + u_{im}] = \pi^2/6 + 1 + \sum_{t=1}^{m-1} D_{i,t}^2 . \tag{10.20}$$

(A different number of terms is needed for each equation to account for the correlations.)

11

Two Part and Sample Selection Models

Two part models describe situations in which the ordered choice is part of a two stage decision process. In a typical situation, an individual decides whether or not to participate in an activity then, if so, decides how much. The first decision is a binary choice. The intensity outcome can be of several types – what interests us here is an ordered choice. In the example below, an individual decides whether or not to be a smoker. The intensity outcome is how much they smoke. The sample selection model is one in which the participation "decision" realtes to whether the data on the outcome variable will be observed, rather than whether the activity is undertaken. This chapter will describe several types of two part and sample selection models

11.1 Inflation Models

Harris and Zhao (2007) analyzed a sample of 28,813 Australian individuals' responses to the question "How often do you now smoke cigarettes, pipes or other tobacco products?" [Data are from the Australian National Drug Strategy Household Survey, NDSHS (2001).] Responses were "zero, low, moderate, high," coded 0,1,2,3. Figure 11.1 below reproduces their Figure 3 (page 1095). The leftmost bar of each set shows the sample histogram. The spike at zero shows a considerable excess of zeros compared to what might be expected in an ordered choice model. The authors reason that there are numerous explanations for a zero response: "genuine nonsmokers, recent quitters, infrequent smokers who are not currently smoking and potential smokers who might smoke when, say, the price falls." It is also possible that the zero response includes some individuals who prefer to identify themselves as nonsmokers. The question is ambiguously worded, but arguably, the group of interest is the genuine nonsmokers. This suggests a type of latent class arrangement in the population. There are (arguably) two types of zeros, the one of interest, and another type generated by the appearance of the respondent in the latent class of people who respond zero when another response would actually be appropriate. The end result is an inflation of the proportion of zero responses in the data. A "Zero Inflation" model is proposed to accommodate this failure of the base case model.

Zero inflation as a formal model to explain data such as these originates in Lambert's (1992) study of quality control in industry. Sampling for defectives in a production process can produce two types of zeros (per unit of time). The process may be under control, or it may be out of control and the observer happens to draw zero defectives in a particular sample. This inflates the number of zeros in a sample beyond what would be expected by a count model such as the Poisson model – the modification named the ZIP (zero inflated) or ZAP (zero altered) Poisson model. [See also Heilbron (1994), Hinde et al. (1998), Mullahy (1997) and Greene (1994).]

Harris and Zhao proposed the following zero inflated ordered probit (ZIOP) model:

Participation equation:

Regime 0 for nonparticipation (nonsmoker), Regime 1 for participation,

$$\begin{aligned} &r^* = \boldsymbol{\alpha}'\mathbf{z} + u,\ u \sim N[0,1], \\ &r = 1 \text{ if } r^* > 0,\ 0 \text{ otherwise}, \\ &\text{Prob}(r = 1|\mathbf{z}) = \Phi(\boldsymbol{\alpha}'\mathbf{z}). \end{aligned} \tag{11.1}$$

Activity equation:

$$y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon,\ \varepsilon \sim N[0,1], \text{ independent of } u,$$
$$y = j \text{ if } \mu_{j-1} < y^* \leq \mu_j,\ j = 0,1,...,J.$$

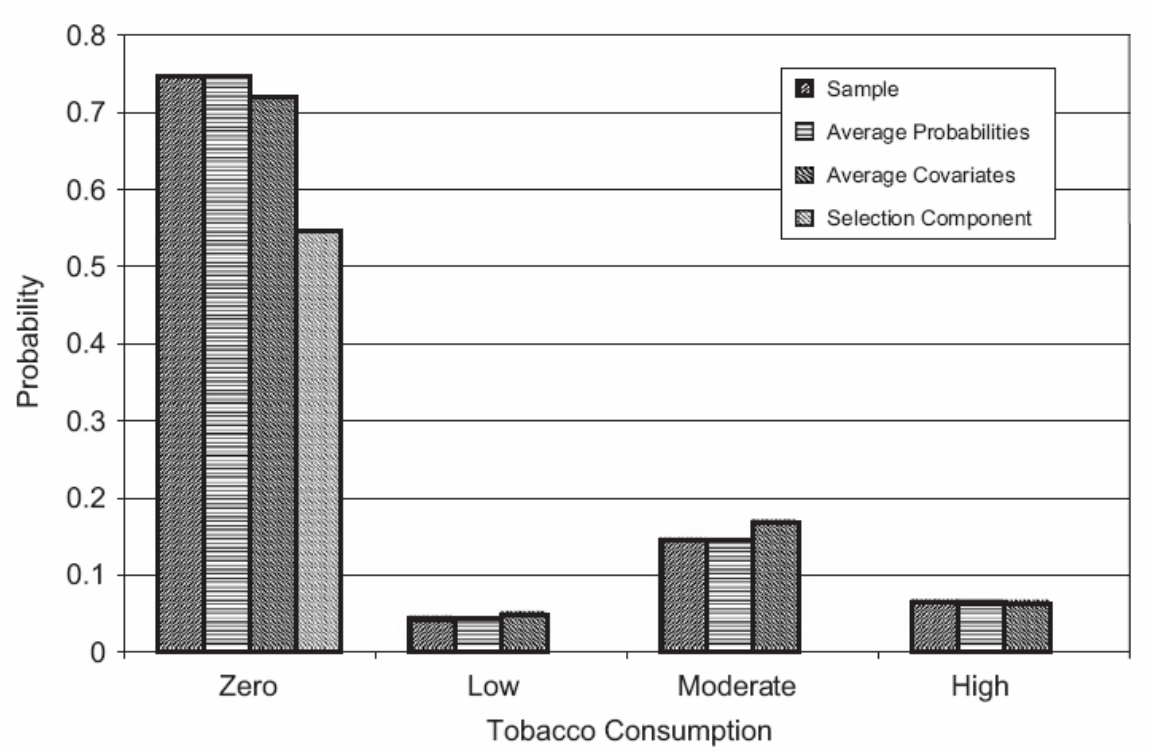


Figure 11.1 Tobacco Consumption Survey and Model Results

(At the risk of some confusion below, we have modified Harris and Zhao's notation to conform to the conventions we have used up to this point.) Thus, a standard probit model governs participation and our familiar ordered probit model governs "true" activity. The observed activity level, however, is not y. It is

$$y_o = r \times y.$$

Nonparticipants and some participants reports zeros. Thus, the zero outcome occurs when $r = 0$ and when $r = 1$ *and* $y = 0$. Therefore, the zero outcome is inflated by the $r = 0$ regime. The applicable probabilities for the observed outcomes are

$$\begin{aligned} \text{Prob}(y_o = 0 \mid \mathbf{x},\mathbf{z}) &= \text{Prob}(r = 0 \mid \mathbf{z}) + \text{Prob}(r = 1 \mid \mathbf{z}) \times \text{Prob}(y = 0 \mid \mathbf{x}, r = 1), \\ \text{Prob}(y_o = j \mid \mathbf{x},\mathbf{z}) &= \text{Prob}(r = 1 \mid \mathbf{z}) \times \text{Prob}(y = j \mid \mathbf{x}, r = 1). \end{aligned} \tag{11.2}$$

Note at this point, by dint of the independence of ε and u, Prob($y = 0 \mid \mathbf{x}$, $r = 1$) = Prob($y = 0 \mid \mathbf{x}$). We will relax this assumption later.

With the assumption of joint normality of ε and u, the associated probabilities are obtained from those of the binary probit model and the ordered probit model;

$$\begin{aligned} \text{Prob}(y_o = 0 \mid \mathbf{x},\mathbf{z}) &= [1 - \Phi(\boldsymbol{\alpha}'\mathbf{z})] + \Phi(\boldsymbol{\alpha}'\mathbf{z}) \times \Phi(0 - \boldsymbol{\beta}'\mathbf{x}), \\ \text{Prob}(y_o = j \mid \mathbf{x},\mathbf{z}) &= \Phi(\boldsymbol{\alpha}'\mathbf{z}) \times [\Phi(\mu_j - \boldsymbol{\beta}'\mathbf{x}) - \Phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x})], j = 1,...,J, \end{aligned} \tag{11.3}$$

with the same normalization as earlier, $\mu_{-1} = -\infty$, $\mu_0 = 0$, $\mu_J = +\infty$. The log likelihood function is built up as the sum of the logs of the probabilities of the observed outcomes.

An extension which would seem to be appropriate for this application is to allow the unobserved effects in the participation equation and the activity equation to be correlated (producing a ZIOPC model). Thus, we now have

$$\begin{pmatrix} u \\ \varepsilon \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

The correlation coefficient, ρ, is now an additional parameter to be estimated. With this modification, we no longer have Prob(y = 0 | **x**, r = 1) = Prob(y = 0 | **x**); the former is now a probability from the bivariate normal distribution. The probabilities of the observed outcomes become

$$\begin{aligned} \text{Prob}(y_o = 0 \mid \mathbf{x},\mathbf{z}) &= [1 - \Phi(\boldsymbol{\alpha}'\mathbf{z})] + \Phi_2(\boldsymbol{\alpha}'\mathbf{z}, -\boldsymbol{\beta}'\mathbf{x}, -\rho), \\ \text{Prob}(y_o = j \mid \mathbf{x},\mathbf{z}) &= \Phi_2(\boldsymbol{\alpha}'\mathbf{z}, \mu_j - \boldsymbol{\beta}'\mathbf{x}, -\rho) - \Phi_2(\boldsymbol{\alpha}'\mathbf{z}, \mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}, -\rho), j = 1,\ldots,J, \end{aligned} \tag{11.4}$$

where $\Phi_2(.,.,.)$ denotes the probability of a joint event from the bivariate normal cdf. This modification drastically alters the partial effects in the model. To organize these in a convenient fashion, we adopt the authors' device. Let $\mathbf{x}^* = (\mathbf{x}_o,\mathbf{x}_c,\mathbf{z}_o)$ so that $\mathbf{x}_o$ is variables in **x** that are not also in **z**, $\mathbf{x}_c$ is variables that are in both **x** and **z**, and $\mathbf{z}_o$ is variables in **z** that are not in **x**. By rearranging and reordering the parameter vectors, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ into $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_o,\boldsymbol{\beta}_c,\mathbf{0})$ and $\boldsymbol{\alpha}^* = (\mathbf{0},\boldsymbol{\alpha}_c,\boldsymbol{\alpha}_o)$, then $\boldsymbol{\beta}'\mathbf{x} = \boldsymbol{\beta}^{*\prime}\mathbf{x}^*$ and $\boldsymbol{\alpha}'\mathbf{z} = \boldsymbol{\alpha}^{*\prime}\mathbf{x}^*$. We can thus obtain the partial effects by differentiating with respect to $\mathbf{x}^*$ and obtaining the needed decomposition. Then, with this in place,

$$\begin{aligned} \frac{\partial \text{Prob}(y_o = 0 \mid \mathbf{x}^*)}{\partial \mathbf{x}^*} &= \left[\Phi\left(\frac{-\boldsymbol{\beta}'\mathbf{x} + \rho\boldsymbol{\alpha}'\mathbf{z}}{\sqrt{1-\rho^2}}\right) - 1\right]\phi(\boldsymbol{\alpha}'\mathbf{z})\boldsymbol{\alpha}^* - \Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z} - \rho\boldsymbol{\beta}'\mathbf{x}}{\sqrt{1-\rho^2}}\right)\phi(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}^*, \\ \frac{\partial \text{Prob}(y_o = j \mid \mathbf{x}^*)}{\partial \mathbf{x}^*} &= \left[\Phi\left(\frac{\mu_j - \boldsymbol{\beta}'\mathbf{x} + \rho\boldsymbol{\alpha}'\mathbf{z}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x} + \rho\boldsymbol{\alpha}'\mathbf{z}}{\sqrt{1-\rho^2}}\right)\right]\phi(\boldsymbol{\alpha}'\mathbf{z})\boldsymbol{\alpha}^* \\ &\quad + \left[\phi(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x})\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z} + \rho(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x})}{\sqrt{1-\rho^2}}\right) - \phi(\mu_j - \boldsymbol{\beta}'\mathbf{x})\Phi\left(\frac{\boldsymbol{\alpha}'\mathbf{z} + \rho(\mu_j - \boldsymbol{\beta}'\mathbf{x})}{\sqrt{1-\rho^2}}\right)\right]\boldsymbol{\beta}^*. \end{aligned} \tag{11.5}$$

These results are likely to bear little resemblance to the raw coefficients, particularly for variables which appear in both equations.

Testing the null hypothesis of the ZIOP model against the alternative of the ZIOPC model is a simple test of the hypothesis that ρ equals zero. This can be done using a Wald (t) test or a likelihood ratio test. Testing for the inflation effects is more complicated however. The obvious restriction, $\boldsymbol{\alpha} = \mathbf{0}$, does not remove the inflation effect; it makes the regime probabilities both equal to one half. What is needed to remove the inflation effect is $\boldsymbol{\alpha}'\mathbf{z} \to \infty$, which cannot be imposed. The hypotheses are not nested. Greene (1994) proposed using the Vuong (1989) test for this hypothesis in the context of the zero inflated Poisson model. Denote the probability for the observed outcome from the inflation model as $f_I(y_o,r|\mathbf{x},\mathbf{z})$ and that for the uninflated model as $f_U(y_o,r|\mathbf{x},\mathbf{z})$. Then,

$$m_i = \log\left(\frac{f_1(y_{o,i}, r_i \mid \mathbf{x}_i, \mathbf{z}_i)}{f_U(y_{o,i}, r_i \mid \mathbf{x}_i, \mathbf{z}_i)}\right). \tag{11.6}$$

The test statistic is

$$V = \frac{\sqrt{N}\ (1/\mathrm{N})\Sigma_{i=1}^{N} m_i}{\sqrt{(1/N)\Sigma_{i=1}^{N}(m_i - \bar{m})^2}} = \frac{\sqrt{N}\ \bar{m}}{s_m}. \tag{11.7}$$

The limiting distribution of V under the null hypothesis of no difference is N(0,1). The test is directional. Large positive values favor the inflation model; large negative values favor the uninflated model. The inconclusive region for a 5% significance level would be (-1.96,+1.96). Given the greater number of parameters in the inflation model, it will be rare for V to be strongly negative. It will often strongly favor the larger model.

Brooks, Harris and Spencer (2007) applied the same style of analysis to the policy decisions of the members of the Bank of England Monetary Policy Committee. In this study, the participation equation is a decision to adjust monetary policy (at all). The activity equation is whether rates should decrease ($y_o = 0$), stay the same ($y_o = 1$) or increase ($y_o = 2$). (The model of Eichengreen, Watson and Grossman (1985) is developed on this logic as well.) In this case, the no change result can occur because of a decision not to change rates, or by an inclination to change rates followed later by a decision not to. Thus, the model produces "ones inflation."

11.2 Sample Selection Models

The familiar sample selection model was extended to binary choice models by Wynand and van Praag (1981) and Boyes, Hoffman and Low (1989). A variety of extensions have also been developed for ordered choice models, both as sample selection (regime) equations and as models for outcomes subject, themselves, to sample selectivity. We consider these two cases and some related extensions.

The models of sample selectivity in this area are built as extensions of Heckman's (1979) canonical model,

$$\begin{aligned}
&\textit{Probit Participation Equation}:\\
&\quad z_i^* = \boldsymbol{\alpha}'\mathbf{w}_i + u_i,\\
&\quad z_i = 1[z_i^* > 0].\\
&\textit{Regression Activity Equation}:\\
&\quad y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\\
&\quad (\varepsilon_i, u_i) \sim N[(0,0),(1,\rho\sigma_\varepsilon,1)].\\
&\textit{Observation:}\ \text{For observations with } z_i = 1,\\
&\quad E[y_i^* \mid \mathbf{x}_i, \mathbf{w}_i, z_i = 1] = \boldsymbol{\beta}'\mathbf{x}_i + (\rho\sigma_\varepsilon)[\phi(\boldsymbol{\alpha}'\mathbf{w}_i)/\ \phi(\boldsymbol{\alpha}'\mathbf{w}_i)]\\
&\qquad\qquad = \boldsymbol{\beta}'\mathbf{x}_i + \theta\lambda_i.
\end{aligned} \tag{11.8}$$

Estimation of the regression equation by least squares while ignoring the selection issue produces biased and inconsistent estimators of all the model parameters. Estimation of this model by two step methods is documented in a voluminous literature, including Heckman (1979) and Greene (2008a). The two step method involves estimating $\boldsymbol{\alpha}$ first in the participation equation using an ordinary probit model, then computing an estimate of λ_i, $\hat{\lambda}_i = \phi\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)/\Phi\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)$, for each individual in the selected sample. At the second step, an estimate of $(\boldsymbol{\beta},\theta)$ is obtained by linear

regression of y_i on $\mathbf{x}_i$ and $\hat{\lambda}_i$. Necessary corections to the estimated standard errors are described in Heckman (1979), Greene (1981,2008b), and, in general terms, in Murphy and Topel (2002).

11.2.1 A Sample Selected Ordered Probit Model

Consider a model of educational attainment or performance in a training or vocational education program (e.g., low, median, high), with selection into the program as an observation mechanism. [Boes (2007) examines a related case, that of a treatment, *D* that acts as an endogenous dummy variable in the ordered outcome model.] The structural equations would be

$$
\begin{aligned}
&\textit{Selection Equation}:\\
&\quad z^* = \boldsymbol{\alpha}'\mathbf{w} + u,\\
&\quad z = 1[z^* > 0].\\
&\textit{Ordered Probit Outcome}:\\
&\quad y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon,\\
&\quad y = j \text{ if } \mu_{j-1} < y^* \leq \mu_j.\\
&\textit{Observation Mechanism}:\\
&\quad y,\mathbf{x} \text{ observed when } z = 1,\\
&\quad (\varepsilon,u) \sim \mathrm{N}[(0,0),(1,\rho,1)].
\end{aligned}
\tag{11.9}
$$

In this situation, the "second step" model is nonlinear. The received literature contains many applications in which authors have "corrected for selectivity" by following the logic of the Heckman two step estimator, that is, by constructing $\lambda_i = \phi(\boldsymbol{\alpha}'\mathbf{w}_i)/\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)$ from an estimate of the probit selection equation and adding it to the outcome equation. [See, e.g., Greene (1994). Several other examples are provided in Greene (2008b).] However, this is only appropriate in the linear model with normally distributed disturbances. An explicit expression, which does not involve an inverse Mills ratio, for the case in which the unconditional regression is $\mathrm{E}[y|\mathbf{x},\varepsilon] = \exp(\boldsymbol{\beta}'\mathbf{x} + \varepsilon)$ is given in Terza (1998). A template for nonlinear single index function models subject to selectivity is developed in Terza (1998) and Greene (2006, 2008a, Sec. 24.5.7). Applications specifically to the Poisson regression appear in several places, including Greene (1995, 2005). The general case typically involves estimation either using simulation or quadrature to eliminate an integral involving *u* in the conditional density for *y*. Cases in which both variables are discrete, however, are somewhat simpler. A near parallel to the model above is the bivariate probit model with selection developed by Boyes, Hoffman and Low (1989) in which the outcome equation above would be replaced with a second probit model. [Wynand and van Praag (1981) proposed the bivariate probit/selection model, but used the two step approach rather than maximum likelihood.] The log likelihood function for the bivariate probit model is given in Boyes et al. (1989) and Greene (2008a, p. 896):

$$
\begin{aligned}
\log L = &\sum_{z=0} \log\Phi(-\boldsymbol{\alpha}'\mathbf{w})\\
&+\sum_{z=1,y=0} \log\Phi_2(-\boldsymbol{\beta}'\mathbf{x},\boldsymbol{\alpha}'\mathbf{w},-\rho)+\sum_{z=1,y=1} \log\Phi_2(\boldsymbol{\beta}'\mathbf{x},\boldsymbol{\alpha}'\mathbf{w},\rho).
\end{aligned}
\tag{11.10}
$$

A straightforward extension of the result provides the log likelihood for the ordered probit case,

$$
\begin{aligned}
\log L = &\sum_{z=0} \log\Phi(-\boldsymbol{\alpha}'\mathbf{w})\\
&+\sum_{z=1} \sum_{j=0}^{J} m_{ij} \log\left[\Phi_2(\mu_j - \boldsymbol{\beta}'\mathbf{x},\boldsymbol{\alpha}'\mathbf{w},\rho) - \Phi_2(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x},\boldsymbol{\alpha}'\mathbf{w},\rho)\right],
\end{aligned}
\tag{11.11}
$$

where $m_{ij} = 1$ if $y_i = j$.

Essentially this model is applied in Popuri and Bhat (2003) to a sample of individuals who chose to telecommute ($z = 1$) or not ($z = 0$) then, for those who do telecommute, the number of days that they do. We note two aspects of this application that do depart subtly the sample selection application: (1) the application would more naturally fall into the category of a hurdle model composed of a participation equation and an activity equation given the decision to participate – in the latter, it is known that the activity level is positive. [See Cragg (1971) and Mullahy (1986).] Thus, unlike the familiar choice case, the zero outcome is not possible here. (2) The application would fit more appropriately into the sample selection or hurdle model frameworks for count data such as the Poisson model. [See, again, Mullahy (1986), Terza (1994), Greene (1995) and Greene (2007a).] Bricka and Bhat (2006) is a similar application applied to a sample of individuals who did (z=1) or did not ($z = 0$) underreport the number of trips in a travel based survey. The activity equation is the number of trips underreported for those who did. This study, like its predecessor could be framed in a hurdle model for counts, rather than an ordered choice model. A third vexing aspect of this type of model emerged here as well. The authors report that the estimated correlation gravitated to +1.000 during estimation, and the log likelihood function continue to increase as it did so. This is to be expected. The ordered choice model can be decomposed into an equivalent set of binary choices, $1(y = 0)$, $1(y > 0)$, $1(y > 1)$, and so on. Thus, it can be seen that the hurdle equation replicates one of the embedded binary choices in the ordered choice model. Because of this redundancy, it is entirely natural that the equations would appear to be perfectly correlated.

Table 11.1 presents estimates of a sample selection model. We have used the choice of *PUBLIC* insurance as the selection mechanism. About 87% of the sample choose the public insurance. We speculate that the factors underlying the motivation to purchase the insurance are also related to the response of health satisfaction. The full model is

$$
\begin{aligned}
&PUBLIC_i^* = \alpha_1 + \alpha_2\, AGE_i + \alpha_3\, EDUC_i + \alpha_4\, HANDDUM_i + u_i,\\
&PUBLIC_i = 1[PUBLIC_i^* > 0],\\
&HEALTH_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\\
&HEALTH_i = j \text{ if } \mu_{j-1} < HEALTH_i^* \leq \mu_j,\\
&(HEALTH_i, \mathbf{x}_i) \text{ observed when } PUBLIC_i = 1,\\
&(u_i.\ \varepsilon_i) \sim N_2[(0,1), (1,1,\rho)],
\end{aligned}
$$

using the same set of regressors as previously. The estimate of ρ suggests that the conjecture might be correct. On average, the factors that motivate insurance purchase seem also to motivate a higher response to the health satisfaction question.

11.2.2 Models of Sample Selection with an Ordered Probit Selection Rule

As noted earlier, the binary probit model is a special case of the ordered probit model. The extension of the sample selection model would follow from replacing the participation equation with

Ordered Probit Participation Equation:

$$
\begin{aligned}
&z_i^* = \boldsymbol{\alpha}'\mathbf{w}_i + u_i, \qquad (11.12)\\
&z_i = j \text{ if } \mu_{j-1} < z_i^* \leq \mu_j.
\end{aligned}
$$

Then, the objective is to recast the conditional mean function, $E[y_i^* \mid \mathbf{x}_i, \mathbf{w}_i, z_i = j]$ and determine an appropriate estimator and set of inference procedures. A typical application (several of those

listed below) considers an "Educational Attainment" participation equation (*secondary*, *college*, *graduate*) and an outcome equation such as an earnings equation.

Table 11.1 Estimated Ordered Probit Sample Selection Model

Binomial Probit Model	
Dependent variable	PUBLIC
Number of observations	4483
Log likelihood function	-1471.427
Restricted log likelihood	-1711.545

Variable	Coeff.	Standard Error	b/St.Er.	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
	Index function for probability				Single Equation probit			
Constant	3.4512	.1622	21.267	.0000	3.5925	.1651	21.758	.0000
AGE	-.0054	.0025	-2.181	.0292	-.0027	.0024	-1.110	.2670
EDUC	-.1804	.0093	-19.394	.0000	-.1967	.0094	-21.016	.0000
HANDDUM	.6710	.0803	8.353	.0000	.2881	.0980	2.939	.0033
	Index function for ordered probit							
Constant	2.2347	.1270	17.590	.0000	Binary Choice Model Predictions			
AGE	-.0160	.0016	-9.780	.0000	Predicted			
EDUC	-.0314	.0092	-3.398	.0007	Actual	0	1	Total
INCOME	.2384	.0994	2.399	.0164	0	164	408	572
MARRIED	-.0093	.0386	-.242	.8089	1	141	3770	3911
KIDS	.0545	.0371	1.466	.1427	Total	305	4178	4483
	Threshold parameters for index							
Mu(1)	.9695	.0394	24.581	.0000				
Mu(2)	2.2399	.0524	42.718	.0000				
Mu(3)	2.7091	.0547	49.519	.0000				
Rho(u,e)	.8080	.0452	17.880	.0000				

Garen (1984) builds directly on the Heckman model. He departs from a model in which

$$\begin{aligned} y_i \mid \mathbf{x}_i, z_i = 0] &= \boldsymbol{\beta}_0'\mathbf{x}_i + \varepsilon_{i0}, \\ y_i \mid \mathbf{x}_i, z_i = 1] &= \boldsymbol{\beta}_1'\mathbf{x}_i + \varepsilon_{i1}, \\ z_i^* &= \boldsymbol{\pi}_1'\mathbf{x}_i + \boldsymbol{\pi}_2'\mathbf{w}_i + u_i,\ z_i = 1[z_i^* > 0], \end{aligned} \tag{11.13}$$

which is similar to the selection model shown above. [As stated, it is a "mover/stayer model." See, e.g., Nakosteen and Zimmer (1980) and Greene (2008a, p. 888).] Garen's suggestion from here suggests how to proceed if z_i is continuous – i.e., if z_i^* were the observation. He proposes to treat z_i as if it were observed in the form of integer values, 1,...,n, noting that the continuous variable emerges as $n \to \infty$. There is, then a different regression equation for each value of z_i. What follows is an analysis of a transformed regression equation that is augmented with powers of z_i and products of z_i and $\mathbf{x}_i$. While not a sample selection treatment as such, this does point in the direction of a formal sample selection treatment based on the ordered probit model.

Terza (1987) develops the two step estimator for a regression model in which one of the regressors is generated by an ordered probit model without regressors. The structural equations are equivalent to

$$\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + \theta q_i + \varepsilon_i, \\ q_i^* &= \alpha + u_i, \\ q_i &= j \text{ if } \mu_{j-1} < q_i^* \leq \mu_j, \\ (\varepsilon_i, u_i) &\sim \mathrm{N}[(0, 0),(\sigma_\varepsilon^2, \rho\sigma_\varepsilon, 1)]. \end{aligned} \tag{11.14}$$

It is convenient to define (once again)

m_{ij} = 1 if $q_i = j$ and m_{ij} = 0 otherwise.

Under these assumptions, Terza's main result is

$$E[y_i|\mathbf{x}_i,m_{i0},m_{i1},...,m_{iJ}] = \boldsymbol{\beta}'\mathbf{x}_i + (\theta\rho)f_i,$$

Where (11.15)

$$f_i = \sum_{j=0}^{J} m_{ij}\left(\frac{\phi(\mu_{j-1}-\alpha)-\phi(\mu_j-\alpha)}{\Phi(\mu_j-\alpha)-\Phi(\mu_{j-1}-\alpha)}\right).$$

[A similar result for the conditional mean of a doubly truncated variable appears in Maddala (1983, p. 366).] Terza goes on to propose a two step estimation procedure. The first step involves maximum likelihood estimation of $(\alpha,\mu_{-1},\mu_0,\mu_1,...,\mu_J)$. This can be done (first noting that as usual, $\mu_{-1} = -\infty$, $\mu_1 = 0$ and $\mu_J = \infty$) using only the sample proportions in the J+1 cells. The model for q_i implies Prob($q_i > 0$) = $\Phi(\alpha)$, so the estimator of α is $\Phi^{-1}(1-P_0)$. Continuing, Prob($q_i > 1$) = $\Phi(\mu_1 - \alpha)$ which suggests a method of moments estimator of μ_1 based on P_1, and so on. With these estimates in hand, he then proposes linear regression of $\mathbf{y}$ on $\mathbf{X}$ and $\hat{\mathbf{f}}$ to estimate $\boldsymbol{\beta}$ and $(\theta\rho)$. (A method of computing appropriate standard errors is presented later.)

As Terza (1987) notes (p. 278) his model is not a correction for selection because the values of the dependent variable are observed for all observations. (The use of the constructed regressor is a means to another end, consistent estimation of $\boldsymbol{\beta}$.) On the other hand, by a minor rearrangement of terms, the results are precisely what is needed for a model of sample selection. First, while retaining the ordered probit observation mechanism for q_i, replace the constant α with the mean of the latent regression, $\boldsymbol{\alpha}'\mathbf{w}_i$. Second, we note that in the "selection on j" case, we observe not $(m_{i0},m_{i1},...,m_{iJ})$ in full, but only one of them. Terza's results then imply

$$E[y_i|\mathbf{x}_i,\mathbf{w}_i,q_i=j] = E[y_i|\mathbf{x}_i,m_{ij}=1] = \boldsymbol{\beta}'\mathbf{x}_i + \gamma\left(\frac{\phi(\mu_{j-1}-\boldsymbol{\alpha}'\mathbf{w}_i)-\phi(\mu_j-\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\mu_j-\boldsymbol{\alpha}'\mathbf{w}_i)-\Phi(\mu_{j-1}-\boldsymbol{\alpha}'\mathbf{w}_i)}\right). \quad (11.16)$$

This is the result needed to complete the sample selection model. The same two step method can now be applied. Terza's method of computing corrected asymptotic standard errors is essentially unchanged.

The model appears in this form in Terza (1983). [See, as well, Vella (1998, p. 148).] Jiminez and Kugler (1987) appears to be the first formal application of the preceding sample selection model. The application is an earnings equation for the Bogota subsample of a 1979-1981 nationwide survey of graduates in Colombia The selection mechanism is determined by participation in a vocational and technical training course (SENA), recorded as *none*, *short* or *long*. The authors derived the conditional mean function from first principles; the derivation follows naturally from earlier results in Maddala (1983), Garen (1984), Heckman (1979), Kenny et al. (1979), Lee and Trost (1978) and Trost and Lee (1978). Kao and Wu (1990) applied the same model to an analysis of bond yields in which the selection mechanism assigns bonds to risk classes by a rating agency. [See, as well, Acharya (1988) for a more elaborate development of the sample selection model.]

Frazis's (1993) study is similar to Jimenez and Kugler. This study analyzes earnings of high school seniors from the National Longitudinal Study of the High School Class of 1972. A panel of seniors was interviewed in 1972, then again five times between 1973 and 1986. Frazis's analysis departed from the basic framework in two ways. The earnings equation is

$$\log y = \boldsymbol{\beta}'\mathbf{x} + \Sigma_j \gamma_j S_j + \delta XS + \phi u + \lambda uS + \varepsilon,$$

where y is earnings, $\mathbf{x}$ is a vector of control variables, S_j is a set of dummy variables that equal one if at least the level of schooling, j, is attained and zero otherwise, XS is interactions of the school attainment dummy variables with X and u represents "aspects of the ability to acquire human capital that are unobservable to the researcher." Thus, since schooling level is the ordered selection mechanism, as stated, this model resembles a treatment effects model, and is also similar to Terza's (1987) formulation. (Motivation for the parts of the equation are given in the paper.) However, note once again, that the observation will be conditioned not on all S_j, but only on the one that corresponds to the individual's schooling level. Estimates of $E[u|S_j]$ to serve as the proxy for u in the earnings equation are obtained by estimating the ordered probit model for schooling level and computing the conditional mean function given earlier. The estimating equation (fit by ordinary least squares) is obtained by replacing u in the equation above (in both places) with

$$\hat{U}_j = \left(\frac{\phi(\mu_{j-1} - \hat{\boldsymbol{\alpha}}'_{j-1}\mathbf{w}_i) - \phi(\mu_j - \hat{\boldsymbol{\alpha}}'_j\mathbf{w}_i)}{\Phi(\mu_j - \hat{\boldsymbol{\alpha}}'_j\mathbf{w}_i) - \Phi(\mu_{j-1} - \hat{\boldsymbol{\alpha}}'_{j-1}\mathbf{w}_i)} \right). \tag{11.17}$$

The second noteworthy point is that, as the author mentions in passing, the ordered probit model provides separate regression coefficients for each level of education. As he notes, this allows negative probabilities. A discussion of aspects of the data set that should prevent this is given.

Two remaining studies of sample selection with ordered probit selection mechanisms are Amel and Liang (1994, 1997) and Butler et al. (1994, 1998). In the first of these, the authors examine firm performance in the banking industry. The conditioning equation used depends on the amount of entry in the market; the authors describe small markets in which entry is described with a simple probit model, and large ones in which ordered probit and truncated Poisson models are used. Butler, Finegan and Siegfried (1998) [see, also Butler et al. (1994)] analyzed performance in economics courses. The selection mechanism is calculus proficiency measured by level of training across several possible courses.

Li and Tobias (2006a) replicated Butler et al. (1998) using a Bayesian method rather than two step least squares. The authors describe an "augmented likelihood function" for the model. With noninformative priors, they "virtually identically" replicated the original results, which suggests that the augmented likelihood function is not equal to the one given above. Technical details are not provided in the paper, but are promised in a no longer existing Iowa State University Economics Department working paper. [Li and Tobias (2006c).] The working paper is reincarnated under the same title in Li and Tobias (2006b). There the authors note that the dependent variable in the regression is actually a grade level, which is also discrete and ordered. The model in (2006b) is a treatment effects model in a triangular system with the outcome of the first ordered probit regression, in the form of a set of endogenous dummy variables, appearing on the right hand side of a second ordered outcome model, the grade attainment,. [Sajaia (2008) is vaguely related to this, however, his treatment of the recursive model builds a simultaneous equations system in the latent regression, which seems difficult to motivate. This paper merely documents a *Stata* program, and does not provide detailed technical background.] The Li and Tobias model without the dummy variables (i.e., under a restriction that their coefficients are zero) would be the bivariate ordered probit model of Section 10.2, so it appears that the authors have rediscovered the MLE for the bivariate model, using a Gibbs sampling and MCMC algorithm rather than classical maximum likelihood. Technical details are omitted from the (2006b) paper, so it is difficult to discern how closely the results resemble each other, but one would expect them, with noninformative priors, to give roughly the same numerical results.

Missing from the preceding and from the received literature is a maximum likelihood estimator for the ordered probit sample selection model. One reason one might wish to consider an MLE as an alternative approach is that the two step estimators do not produce an estimator of ρ, which is likely to be an interesting parameter, for example, if one wished to test for "selectivity." [In the basic case, there is a method of moments estimator of ρ available – See Heckman (1979) and Greene (1981). However, none has been derived for the ordered choice case. An analog to the estimator developed by Heckman (1979) would be straightforward. However, it will have the same shortcoming as the one in the basic model. As shown in Greene (1981), the estimator is not bounded by -1 and +1. Moreover, even when it does fall in the right range, no inference is possible. (This latter point is of minor consequence. In the original model above, inference is possible about $\theta = \rho\sigma$ based on the OLS results, and $\rho = 0$ is both necessary and sufficient for $\theta = 0$, as σ cannot be zero in a sensible model.)

The log likelihood for the original sample selection model (binary selection and linear regression) is given in Greene (2008a, eq. 24-33) and in Econometric Software (2007);

$$\log L = \left\{ \begin{array}{l} \sum_{z_i=0} \log \Phi(-\boldsymbol{\alpha}'\mathbf{w}_i) + \\ \sum_{z_i=1} \log\left[\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma_\varepsilon}\right)\Phi\left(\frac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon + \boldsymbol{\alpha}'\mathbf{w}_i}{\sqrt{1-\rho^1}}\right)\right] \end{array} \right\}. \tag{11.18}$$

This estimator, though apparently much less frequently used than the two step method, is available as a preprogrammed procedure in contemporary software such as *Stata* and *NLOGIT*. Note that it is a full information maximum likelihood estimator for all the parameters in the model. The estimator is not less robust than the two step estimator; both are fully parametric based on the bivariate normal distribution.

The counterpart for an ordered probit sample selection model will replace the term Φ(.) in the square brackets with

$$F_i = \begin{array}{l} \Phi\left(\frac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon - (\mu_j - \boldsymbol{\alpha}'\mathbf{w}_i)}{\sqrt{1-\rho^2}}\right) - \\ \Phi\left(\frac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon - (\mu_{j-1} - \boldsymbol{\alpha}'\mathbf{w}_i)}{\sqrt{1-\rho^2}}\right) \end{array}, \tag{11.19}$$

and the term $\Phi(-\boldsymbol{\alpha}'\mathbf{w}_i)$ with

$$\text{Prob}(z_i \neq j \mid \mathbf{w}_i) = 1 - \left[\Phi\left(\mu_j - \boldsymbol{\alpha}'\mathbf{w}_i\right) - \Phi\left(\mu_{j-1} - \boldsymbol{\alpha}'\mathbf{w}_i\right)\right]. \tag{11.20}$$

As stated, this is a conventional maximum likelihood estimator that produces the familiar properties consistency, asymptotic normality, etc. If the selection is "selection on a particular *j*," however, then no more than one of the threshold parameters will be estimable. Assuming that $\boldsymbol{\alpha}$ contains a constant term, if selection is on $j = 0$, then the second probability becomes zero and μ_0 already equals zero. If selection is on $j = 1$, then μ_0 in the second probability is zero and the constant in $\boldsymbol{\alpha}$ is identified, while in the first probability, μ_1 is estimable distinct from the constant in $\boldsymbol{\alpha}$. If selection is on $j > 1$, then the two probabilities have separate constant terms, but only two

distinct constant terms are estimable. The first constant term estimates (α_0 - μ_j) and the second estimates (α_0 - μ_{j-1}).

Full information maximum likelihood based on the probabilities shown above should be a conventional, relatively straightforward exercise. However, there is a simplification that might prove useful. This (and the original model) is an ideal setting to employ the Murphy and Topel (2002) kind of two step estimator. As already seen, we can estimate the ordered probit model in isolation, using maximum likelihood. Let

$$\hat{A}_j = \hat{\mu}_j - \hat{\boldsymbol{\alpha}}'z_i,\ \hat{A}_{j-1} = \hat{\mu}_{j-1} - \hat{\boldsymbol{\alpha}}'z_i.$$

Then, a two step approach can be used in which the log likelihood function maximized at the second step is

$$\begin{aligned}\log L = &\sum_{z_i \neq j} \log\left[1 - (\Phi(\hat{A}_j) - \Phi(\hat{A}_{j-1}))\right] \\ &+ \sum_{z_i = j} \log\left[\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma_\varepsilon}\right)\left\{\begin{array}{l}\Phi\left(\dfrac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon - \hat{A}_j}{\sqrt{1-\rho^2}}\right) - \\ \Phi\left(\dfrac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon - \hat{A}_{j-1}}{\sqrt{1-\rho^2}}\right)\end{array}\right\}\right].\end{aligned} \tag{11.21}$$

Note that the first term is now an irrelevant constant, and the log likelihood function to be maximized is based only on the selected sample. This can be made even more convenient by reparameterizing it with the Olsen (1978) reparameterization, $\theta = 1/\sigma_\varepsilon$ and $\boldsymbol{\gamma} = (1/\sigma_\varepsilon)\boldsymbol{\beta}$. Now, the relevant log likelihood is

$$\log L^* = \sum_{z_i = j} \log\left[\theta\phi(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i)\left\{\begin{array}{l}\Phi\left(\dfrac{\rho(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i) - \hat{A}_j}{\sqrt{1-\rho^2}}\right) - \\ \Phi\left(\dfrac{\rho(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i) - \hat{A}_{j-1}}{\sqrt{1-\rho^2}}\right)\end{array}\right\}\right]. \tag{11.22}$$

Finally, let $\tau = \rho/\sqrt{1-\rho^2}$. Then, the log likelihood simplifies a bit more to

$$\begin{aligned}\log L^* = &\sum_{z_i = j} \log\theta\phi(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i) + \\ &\sum_{z_i = j} \log\left\{\begin{array}{l}\Phi\left(\tau(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i) - \sqrt{1+\tau^2}\,\hat{A}_j\right) - \\ \Phi\left(\tau(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i) - \sqrt{1+\tau^2}\,\hat{A}_{j-1}\right)\end{array}\right\}.\end{aligned} \tag{11.23}$$

Once estimates of θ, $\boldsymbol{\gamma}$ and τ are in hand, estimates of the structural parameters, σ_ε, $\boldsymbol{\beta}$ and ρ, can be obtained by inverting the transformations. This appproach has an additional benefit in that the range of τ is unrestricted, while that of ρ must be restricted to (-1,+1) during estimation.

11.2.3 A Sample Selected Bivariate Ordered Probit Model

Bhat and Singh (2000) extended the preceding methodology to a sample selected bivariate ordered choice model. In their study, the selection mechanism is a multinomial (three outcome) logit model for travel mode choice linked to a bivariate ordered choice model for counts of two travel related activities. To develop the approach, we begin with a simpler case, with a binary selection equation. For two outcomes and a binary regime selection, we would have

Selection Equation:

$$z^* = \boldsymbol{\alpha}'\mathbf{w} + u,$$
$$z = 1[z^* > 0].$$

Bivariate Ordered Probit Activity Equation: (11.24)

$$y_{i,1}^* = \boldsymbol{\beta}_1'\mathbf{x}_{i,1} + \varepsilon_{i,1}, y_{i,1} = j \text{ if } \mu_{j-1} < y_{i,1}^* < \mu_j, j = 0,\ldots,J_1,$$
$$y_{i,2}^* = \boldsymbol{\beta}_2'\mathbf{x}_{i,2} + \varepsilon_{i,2}, y_{i,2} = j \text{ if } \delta_{j-1} < y_{i,2}^* < \delta_j, j = 0,\ldots,J_2,$$

$$\begin{pmatrix} u_i \\ \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta_1 & \theta_2 \\ \theta_1 & 1 & \rho \\ \theta_2 & \rho & 1 \end{pmatrix}\right] = N[\mathbf{0}, \boldsymbol{\Sigma}].$$

As before, the ordered choice outcomes, $(y_{i,1},y_{i,2})$ are observed only when z_i equals one. The joint probability for $z = 1$, $y_{i,1} = j$ and $y_{i,2} = k$ is

$$\begin{aligned} L_i^1 &= \text{Prob}(z_i = 1, y_{i,1} = j, y_{i,2} = k \mid \mathbf{w}_i, \mathbf{x}_{i,1}, \mathbf{x}_{i,2}) \\ &= \begin{bmatrix} \Phi_3[\boldsymbol{\alpha}'\mathbf{w}_i, (\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}), \boldsymbol{\Sigma}] \\ -\Phi_3[\boldsymbol{\alpha}'\mathbf{w}_i, (\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}), (\delta_k - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}), \boldsymbol{\Sigma}] \end{bmatrix} - \\ &\quad \begin{bmatrix} \Phi_3[\boldsymbol{\alpha}'\mathbf{w}_i, (\mu_j - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}), \boldsymbol{\Sigma}] \\ -\Phi_3[\boldsymbol{\alpha}'\mathbf{w}_i, (\mu_{j-1} - \boldsymbol{\beta}_1'\mathbf{x}_{i,1}), (\delta_{k-1} - \boldsymbol{\beta}_2'\mathbf{x}_{i,2}), \boldsymbol{\Sigma}] \end{bmatrix}, \end{aligned} \quad (11.25)$$

where $\Phi_3(\ldots)$ denotes the trivariate normal cdf using mean vector zero and covariance matrix $\boldsymbol{\Sigma}$. The contribution to the likelihood function for observations with $z_i = 0$ is

$$L_i^0 = \text{Prob}[z_i = 0|\mathbf{w}_i] = \Phi(-\boldsymbol{\alpha}'\mathbf{w}_i). \quad (11.26)$$

The log likelihood is

$$\log L = \sum_{z_i=1} \log L_i^1 + \sum_{z_i=0} \log L_i^0. \quad (11.27)$$

The log likelihood requires evaluation of trivariate normal integrals, for which quadrature based methods are available (e.g., Genz (2008) or Drezner (1994)). Alternatively, the GHK simulator may be used.

Bhat and Singh's application replaces the selection equation with a multinomial logit model for mode choice among Q alternatives,

$$m_{i,q}^* = \gamma_q'\mathbf{h}_{i,q} + v_{i,q}, q = 1,\ldots,Q, \quad (11.28)$$

where $v_{i,q}$ has type 1 extreme value (Gumbel) distribution, all independent. Mode choice is then reparameterized as three binary choice equations,

$$M_{i,q}^* = \boldsymbol{\gamma}_q'\mathbf{h}_{i,q} - v_{i,q},\ M_{i,q} = 1 \text{ if } M_{i,q}^* > 0.$$
$$\text{Prob}(M_{iq} = 1|\mathbf{h}_{iq}) = F(\boldsymbol{\gamma}_q'\mathbf{h}_{i,q}) = \frac{\exp(\gamma_q' h_{i,q})}{\Sigma_{s=1}^{Q}\exp(\gamma_s' h_{i,s})},\ q = 1,\ldots,Q. \tag{11.29}$$

The transformation to normality uses Lee's (1983) copula function approach,

$$v_{i,q}^* = \Phi^{-1}[F(\boldsymbol{\gamma}_q'\mathbf{h}_{i,q})]. \tag{11.30}$$

Then, the selection is based on the mode choice, and a separate bivariate ordered probit model applies to each mode. In addition to the choice model parameters that are estimated, there is a set of thresholds, $\mu_{q,1}$ and $\mu_{q,2}$, a set of slopes, $\boldsymbol{\beta}_{1,q}$ and $\boldsymbol{\beta}_{2,q}$, and a set of three correlation coefficients, ρ_q, $\theta_{1,q}$ and $\theta_{2,q}$. In their application, three travel modes were drive alone, shared ride and transit. Two activities are evening commute stops and post home-arrival stops. The full set of parameters in the multinomial logit model and in the set of bivariate ordered probit models are then estimated simultaneously by full information maximum likleihood (subject to a large number of zero restrictions on the parameters of the utility functions and in $\boldsymbol{\Sigma}_q$.) Readers are referred to Bhat and Singh (2000) for details.

11.3 An Ordered Probit Model with Endogenous Treatment Effects

Munkin and Trivedi (2008) have analyzed a model that bears some connection to the selection model proposed in the previous section. The model extension considered involves a set of endogenous "treatment dummy variables." That is,

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\delta}'\mathbf{d}_i + \varepsilon_i,$$
$$y_i = j \text{ if } \mu_{j-1} < y_i^* \leq \mu_j,$$

where y_i is a measure of medical service utilization (actually a count with excess zeros – the ordered choice model is used as an approximation). The additional vector of covariates, $\mathbf{d}_i$, is a set of dummy variables that is the outcome of a choice of treatments; one of M treatments is chosen and for that choice, $d_{im} = 1$ and $d_{im'} = 0$ for all others. (We have included all M treatments in $\mathbf{d}_i$ for pedagogical convenience. In their analysis, one of the dummy variables is immediately dropped from the model since only M-1 are needed to determine the observed outcome.) The treatment outcome is determined by a multinomial probit model of underlying utility across the choices. [See Train (2003) and the large number of sources cited by Munkin and Trivedi for discussion of the multinomial probit model.] The endogeneity of the treatment effects follows from the correlations between the random elements of the random utility equations in the choice model and the random term, ε_i in the ordered choice model. A Bayesian (MCMC) treatment is used to estimate the posterior means of the parameters. A method for estimating models with more general forms of endogenous right hand side variables is suggested in Kawakatsu and Largey (2009).

12

Semiparametric and Nonparametric Estimators and Analyses

The foregoing has surveyed nearly all of the literature on ordered choice modeling. We have, of course, listed only a small fraction of the received applications. But, the full range of methodological developments has been presented, with a single remaining exception. As in many other areas of econometrics, a thread of the contemporary literature has explored the boundaries of the model that are circumscribed by the distributional assumptions. We have limited ourselves to ordered logit and probit models, while relaxing certain assumptions such as homoscedasticity, all within the boundaries of the parametric model. The last strand of literature to be examined is the development of estimators that extend beyond the parametric distributional assumptions. It is useful to organize the overview around a few features of the model, scaling, the distribution of the disturbance, the functional form of the regression, and so on. In each of these cases, we can focus on applications that broaden the reach of the ordered choice model to less tightly specified settings.

There is a long, rich history of semiparametric and nonparametric analysis of binary choice modeling (far too long and rich to examine in depth in this already long survey) that begins in the 1970s, only a few years after analysis of individual binary data became a standard technique. The binary choice literature has two focal points, maximum score estimation [Manski (1975, 1985), Manski and Thompson (1985) and Horowitz (1992)] and the Klein and Spady (1993) kernel based semiparametric estimator for binary choice. (As noted, there is a huge number of other papers on the subject. We are making no attempt to survey this literature.) Some of the more recent developments build on these two (mainly on the second; MSCORE remains to provide a platform for analysis of ordered choices). Surprisingly, the formal extension of the binary choice models to what would seem to be the natural next step, ordered choice, takes place entirely since 2000.

To a very small extent, some of the developments already mentioned move the analysis in the direction of a semiparametric approach. Agresti (1999), for example, notes the extension of GEE methods [see Diggle, Liang and Zeger (1994)] to the ordered choice model. GEE modeling is based more strongly on conditional means and variances than on distributions, and can be viewed as a small step away from the maximum likelihood estimator. (The step is quite small; the formal distributional model is still assumed. One might surmise, however, that the GEE estimator has at least the potential to be robust to failures of the distributional assumption. This remains to be verified, however.) On the other hand, if the latent class model (LCM) that we examined in Section 8.2 is simply interpreted as a mixing model rather than as a latent grouping model, then the LCM certainly qualifies as a semiparametric approach. [See Heckman and Singer (1984) for example.] Likewise, the mixed (random parameters) ordered probit model can also be viewed as a semiparametric estimator; a continuous mixture of underlying distributions that does not adhere to a strict distributional assumption. [See, e.g., McFadden and Train (2001) for discussion of using continuous mixture models to approximate any underlying distribution.] (For the ordered choice model, to achieve full generality in this interpretation, we would want to allow the thresholds, as well as the regression slopes, to be random.)

The received literature on semiparametric (and semi-nonparametric and nonparametric) analysis of ordered choice models is fairly compact. We begin with a study by Chen and Khan (2003) that considers the ordered probit model in the presence of unknown (and not parameterized) heteroscedasticity. Lewbel (2000) goes a step beyond Chen and Khan in allowing the distribution to be unspecified as well. We will then examine Stewart's (2003) parameterized

model that approximates an unknown distibution. Some general observations are collected in Section 12.5. This is not a complete enumeration of this thread of literature (though it is fairly close). Two studies not examined in detail below, but mentioned here are Coppejans (2007) and Klein and Sherman (2002), both of which develop consistent parameter estimators, but are, at the same time, focused somewhat more heavily on methodological aspects of estimation than the papers examined below.

12.1 Heteroscedasticity.

Chen and Khan (2003) propose a semiparametric estimator for the heteroscedastic *ordered probit* model,

$$\begin{aligned} &y_i^* = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \sigma(\mathbf{x}_i)\varepsilon_i, \\ &y_i \;\; = j \;\text{ if }\; \mu_{j-1} < y_i^* \;\leq\; \mu_j, j = 0,1,\ldots,J, \end{aligned} \tag{12.1}$$

(we are adapting their application to our notation – theirs differs in several ways likely to produce ambiguities in the presentation). The issue is whether it is possible efficiently (by semiparametric standards) to estimate $\boldsymbol{\beta}$. Several normalizations are necessary to begin. As usual, $\mu_{-1} = -\infty$ and $\mu_J = +\infty$. Since there is assumed to be a nonzero constant term, α, $\mu_0 = 0$. They restrict attention to the case $J = 2$ (three possible outcomes). "As is always the case with discrete response models, location and scale normalizations are required. As a location normalization, to identify the intercept term, [α] we set [μ_0] = 0. As a scale normalization, we set [μ_1] = 1." (Again, our notation.) The last assumption is, of course, crucial. Heretofore, we have achieved scale normalization by assuming $\sigma_\varepsilon = 1$. The implication of the new assumption in the three outcome is as follows:

$$\begin{aligned} \text{Prob}(y_i = 0 \mid \mathbf{x}_i) \;&=\; \Phi\left(\frac{-\alpha - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma(\mathbf{x}_i)}\right) = P_{0i}, \\ 1 - \text{Prob}(y_i = 2 \mid \mathbf{x}_i) \;&=\; \Phi\left(\frac{1 - \alpha - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma(\mathbf{x}_i)}\right) = 1 - P_{2i}. \end{aligned} \tag{12.2}$$

It follows that

$$\Phi^{-1}(1 - P_{2i}) - \Phi^{-1}(P_{0i}) = \Phi^{-1}(P_{0i} + P_{1i}) - \Phi(P_{0i}) = \frac{1}{\sigma(\mathbf{x}_i)} > 0. \tag{12.3}$$

This suggests that the variance is estimable. The authors propose a kernel estimator,

$$\hat{P}_{ji} = \frac{\frac{1}{H_N}\sum_{l \neq i, l=1}^{N} m_{ij} K\left[\frac{x_l - x_i}{H_N}\right]}{\frac{1}{H_N}\sum_{l \neq i, l=1}^{N} \left[\frac{x_l - x_i}{H_N}\right]}, j = 0, 2. \tag{12.4}$$

(This is a multivariate kernel in any realistic case. For the application, the authors used a product of Epanechnikov kernel functions. Details on selection of the bandwidth may be found in their paper.) With these estimates of P_{i0} and P_{i2} in hand, the estimator of $\sigma(\mathbf{x}_i)$ is

$$\hat{\sigma}(\mathbf{x}_i) = \frac{1}{\Phi^{-1}\left(1-\hat{P}_{i2}\right) - \Phi^{-1}\left(\hat{P}_{i0}\right)}. \tag{12.5}$$

The second step MLEs of α and **β** are obtained by maximizing a log likelihood function,

$$\log \hat{L} = \sum_{i=1}^{N} \tau(\mathbf{x}_i) \log \left[\begin{array}{c} m_{i0}\Phi\left(\dfrac{-\alpha-\boldsymbol{\beta}'\mathbf{x}_i}{\hat{\sigma}(\mathbf{x}_i)}\right) + \\ m_{i1}\left(\Phi\left(\dfrac{1-\alpha-\boldsymbol{\beta}'\mathbf{x}_i}{\hat{\sigma}(\mathbf{x}_i)}\right) - \Phi\left(\dfrac{-\alpha-\boldsymbol{\beta}'\mathbf{x}_i}{\hat{\sigma}(\mathbf{x}_i)}\right)\right) + \\ m_{i2}\Phi\left(\dfrac{\alpha+\boldsymbol{\beta}'\mathbf{x}_i-1}{\hat{\sigma}(\mathbf{x}_i)}\right) \end{array} \right]. \tag{12.6}$$

Where $\tau(\mathbf{x}_i)$ is a trimming function "often adopted in two-step estimators, whose support is assumed to be a compact subset of the support of $\mathbf{x}_i$. For the study done here, $\tau(\mathbf{x}) = 0$ if either predicted probability is outside [0.005,0.995] and 1 otherwise. The Monte Carlo study that follows agrees with expectations; when the ordered probit model is well specified, it performs well, and when it is not, it performs poorly. Likewise confirming expectations, the authors find that when there is pronounced heteroscedasticty, their estimator outperforms the MLE that assumes homoscedastic disturbances.

12.2 A Distribution Free Estimator with Unknown Heteroscedasticty

Lewbel's (2000) formulation of an *ordered choice* model that allows heteroscedasticity of unknown form is

$$\begin{aligned} y_i^* &= z_i + \boldsymbol{\beta}'\mathbf{x}_i + \sigma_i\varepsilon_i, \\ y_i &= j \text{ if } \mu_{j-1} < y_i^* \leq \mu_j, j = 0,1,\ldots,J. \end{aligned}$$

(We rely heavily on Stewart's (2005) very concise exposition of this model.) In this instance, the normalization is transferred to one of the slope coefficients. Lewbel's model is initially formulated in terms of a constant σ, but it is noted that the estimator is robust to heteroscedasticity of unknown form. It is convenient to carry the more general form above. Lewbel's estimator is noniterative and requires only ordinary least squares regressions. The "special variable," z_i whose coefficient is normalized, is required to satisfy certain requirements [see Lewbel (2000) and Stewart (2005).] Among other features, the sign of z_i must be observed. Then, define the indicator $y_{.i} = y_i/J$; values range from 0 to 1. Then construct,

$$\begin{aligned} \tilde{y}_{\cdot i} &= \frac{y_{.i} - 1[z_i > 0]}{f(z_i \mid \mathbf{x}_i)}, \\ \tilde{y}_{ji} &= \frac{1[y_i > j] - 1[z_i > 0]}{f(z_i \mid \mathbf{x}_i)}. \end{aligned} \tag{12.7}$$

The numerators are trivial to compute, however, the density of z given $\mathbf{x}$ requires some additional computation. Stewart (2005, p. 559) navigates some of the developments in the literature for this computation. Assuming the estimator of $f(z|\mathbf{x})$ is in hand and in the estimator of $\tilde{y}_{\cdot i}$, the estimate

of $\boldsymbol{\beta}$ is obtained by least squares regression of $\tilde{y}_{\cdot i}$ on $\mathbf{x}_i$. The estimates of the threshold parameters are the negatives of the constant terms in the J-1 regressions of $\tilde{y}_{ji}$ on $\mathbf{x}_i$.

Lewbel provides this approach for binary, ordered and unordered choice models, censored regressions, and a variety of other settings. Stewart notes, that he found no empirical applications of the ordered choice model, and only a few about binary responses. There have also been "few" studies that compare the estimator to other semiparametric approaches. Little is known about the behavior of this estimator beyond the asymptotic properties that Lewbel, himself, has established in a series of papers [e.g., Lewbel (1997, 2000), Lewbel and Schennach (2007), Honore and Lewbel (2002).]

12.3 A Semi-nonparametric Approach

Stewart (2003, 2005) proposes a model that nests the ordered probit model in a general estimator of an unknown density. The alternative density, proposed by Gallant and Nychka (1987) is

$$f_K(\varepsilon) = \frac{1}{\theta}\left(\sum_{k=0}^{K}\gamma_k\varepsilon^k\right)^2 \phi(\varepsilon). \tag{12.8}$$

The constant, θ, normalizes the density so that it integrates to 1;

$$\theta = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{K}\gamma_k\varepsilon^k\right)^2 \phi(\varepsilon)d\varepsilon\,. \tag{12.9}$$

With the normalization, the density is homogeneous of degree zero in $\boldsymbol{\gamma} = (\gamma_0,\ldots,\gamma_K)$, so the normalization $\gamma_1 = 0$ is imposed. If the remaining $\gamma_k = 0$, the normal distribution results. The class of distributions is defined by the order of the polynomial, K. The model shares a feature with the latent class model examined in Section 8.2; the index, K, is not parametric, and must be located by a specification search. Surprisingly, it turns out that the normal model emerges with $K = 1$ and $K = 2$ as well as $K = 0$; the first model in the series that extends the ordered probit model has $K = 3$. The model selection problem is a bit more straightforward here in that the order of the model is reduced by one if $\gamma_K = 0$, so a likelihood based approach can be used for the specification search.

Stewart notes that the implicit scaling is needed to interpret the coefficients in any ordered choice model. For the application he considers, he suggests that ratios of coefficients are likely to be useful for several reasons. Figure 12.1 is extracted from Table 1 in Stewart (2005). (An alternative model formulation has been omitted.) The OP and SNP estimates are broadly similar, but the least squares estimates show some pronounced differences from both of the others. The SNP model is a parametric extension of the ordered probit model – hence the name "semi-nonparametric." It is not in the same class as the Lewbel or Chen and Khan specifications. The likelihood ratio test rejects the ordered probit model. The results in Figure 12.1 do not include the polynomial parameters or the threshold parameters from the ordered choice models. Figure 12.2 is Table 2 from Stewart's earlier study using the same data and a much larger model. Moving across the results, we see the changes from K=2 (OP) to the 3 and 5 order polynomials. The hypothesis tests against the null model reject the ordered probit model in both cases. The third order model is also rejected in favor of the fifth order one.

Table 1
Job satisfaction model—alternative estimators

	Mean	OP	LLS	SNP(3)
log(earnings)	6.66	1	1	1
log('comparison')	6.66	−2.73 (0.98)	−3.24 (1.49)	−2.86 (0.40)
log(hours)	4.95	−1.66 (0.78)	0.20 (3.38)	−1.22 (0.48)
Male	0.50	−1.56 (0.79)	0.56 (0.90)	−1.25 (0.37)
Age/10	3.72	−1.81 (1.12)	−0.51 (2.28)	−1.35 (0.75)
$Age^2/100$	15.19	0.33 (0.18)	0.03 (0.28)	0.26 (0.10)
Health	0.18	−2.28 (1.11)	−1.27 (0.51)	−1.85 (0.40)
Second job	0.10	−0.91 (0.66)	1.33 (1.53)	−0.85 (0.44)
Temporary	0.06	−1.44 (0.90)	0.31 (1.00)	−1.16 (0.58)
Manager	0.38	1.68 (0.86)	1.63 (1.03)	1.29 (0.33)
Log-likelihood		−6210.14		−6204.58
L-R test of OP				11.13
[*p*-value]				[0.001]

Notes:
(1) Standard errors in parentheses.
(2) Estimators: OP = Ordered Probit estimator, LLS = Lewbel least squares estimator, SNP = Semi- nonparametric estimator,

Figure 12.1 Table 1 From Stewart (2005)

	OP	SNP(3)	SNP(5)
	coef (s.e.)	coef (s.e.)	coef (s.e.)
log(earnings)	0.134 (.054)	0.096 (.050)	0.087 (.056)
log(comp. earn.)	-0.283 (.064)	-0.254 (.060)	-0.323 (.068)
Male	-0.156 (.044)	-0.147 (.044)	-0.130 (.046)
Age/10	-0.210 (.100)	-0.154 (.092)	-0.141 (.104)
$Age^2/100$	0.038 (.013)	0.031 (.011)	0.030 (.013)
Thresholds:			
1	-4.125 (.377)	-4.125	-4.125
2	-3.917 (.376)	-3.879 (.043)	-3.847 (.047)
3	-3.562 (.375)	-3.476 (.093)	-3.390 (.075)
4	-2.995 (.375)	-2.877 (.161)	-2.603 (.125)
5	-2.416 (.374)	-2.319 (.211)	-1.889 (.164)
6	-1.664 (.374)	-1.657 (.254)	-1.151 (.177)
Polynomial:			
1		-0.050 (.193)	0.415 (.062)
2		-0.097 (.088)	0.366 (.252)
3		-0.051 (.021)	-0.073 (.171)
4			-0.087 (.034)
5			-0.002 (.016)
Log-likelihood	-6174.25	-6169.87	-6165.48
Standard deviation	1	0.979	1.369
Skewness	0	0.034	0.064
Kurtosis	3	4.600	4.665
Test sum=0 $[\chi^2(1)]$	7.40	9.26	15.83
p-value	[0.007]	[0.002]	[0.000]

Notes: (1) Sample size = 3895. (2) Models also contain 34 other variables.

[2]The other variables included in the model are as in Table 3 of Clark and Oswald (1996). The data used here are from the current release of Wave 1, while they used the original release. The ordered probit results are very close to theirs, but not identical.

Figure 12.2 Job Satisfaction Application, Extended

12.4 A Partially Linear Model

Bellemare, Melenberg and van Soest (2002) propose the following ordered choice model based on a partially linear (semiparametric) latent regression, ordered probit model:

$$\begin{aligned} y_i^* &= g(\mathbf{z}_i) + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \\ y_i &= j \text{ if } \mu_{j-1} < y_i^* < \mu_j. \end{aligned} \tag{12.10}$$

Their model specifies $\varepsilon_i \sim N[0,\sigma^2]$, however, σ remains unidentified. The usual normalizations of the threshold parameters are also required. There is an interesting intersection of the different aspects of "semiparametric" at this point. It seems that concern about the distribution of ε_i would be a moot point here; if $g(.)$ is unspecified, then it seems unlikely that an observed sample could support estimation of a model that is also built around an unspecified density for ε. The implied probabilities for the model are

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{z}_i) = \Phi\left(\mu_j - g(\mathbf{z}_i) - \boldsymbol{\beta}'\mathbf{x}_i\right) - \Phi\left(\mu_{j-1} - g(\mathbf{z}_i) - \boldsymbol{\beta}'\mathbf{x}_i\right). \tag{12.11}$$

Estimation of the model is suggested using a technique by Hardle, Huet, Mammen and Sperlich (2004) and Severini and Staniswalis (1994). This involves iterating back and forth between maximum likelihood estimation of $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\mu},\sigma)$ conditioned on estimates of $g(\mathbf{z}_i)$ and estimates of $g(\mathbf{z}_i)$ given the other parameters. The former uses the conventional MLE carrying the current estimates of $g(\mathbf{z}_i)$ as known constants. The latter is accomplished by maximizing a separated weighted likelihood function for each i to obtain the current estimate of $g(\mathbf{z}_i)$.

12.5 Semiparametric Analysis

We have examined most of the received developments in the area of semiparametric and nonparametric analyses of the ordered choice model. The central focus of the developments is consistent estimation of the regression slope parameters, $\boldsymbol{\beta}$ in the absence of an assumption about the distribution or the variance of the disturbance. As we have observed repeatedly in the preceding analyses, however, these elements of the model are crucial for translating the coefficient estimates into meaningful characterizations of the underlying data generating process, and these features are absent by design from the semiparametric estimators. Perhaps the signature feature of the ordered choice model is the vexing result that neither the sign nor the magnitude of $\boldsymbol{\beta}$ is informative about the impact of interesting right hand variables on the process that generates the outcome variable. For example, Coppejans (2007) comments at length on the difference in magnitude of a particular coefficient (a fee elasticity) estimated by the ordered probit MLE compared to that obtained by a distribution-free sieve estimator. But, the difference in magnitude observed there is comparable to the difference that would emerge in the same context if he had used an ordered logit model compared to an ordered probit model. The fact that the scaling induced by the distributional model has been obscured in the estimation process is crucial to the finding. That is, the comparison of the estimates of -0.20 for an ordered probit model to a -0.063 for the semiparametric estimator is meaningless without information on the scaling induced by the underlying distributions. No evidence is available to eliminate the possibility that the partial effect in the ordered probit model is actually larger, not smaller, than that in the semiparametric model.

The presence of unaccounted for heteroscedasticity makes this worse. In the Chen and Khan (2003) model, the heteroscedasticity involves the same $\mathbf{x}$ as the mean of the regression. The upshot is that in neither model is $\boldsymbol{\beta}$ the partial effect of interest – indeed, the sign of that

partial effect could be different from that of $\boldsymbol{\beta}$ in all cells of the outcome, since the mean effect, $\boldsymbol{\beta}$, and the variance effect, $\sigma(\mathbf{x})$ would typically have opposite signs. In their formulation of the model, any partial effect will have to include $\partial\sigma(\mathbf{x})/\partial\mathbf{x}$, however, $\sigma(\mathbf{x})$ is not estimated parametrically; we have no idea what this derivative looks like.

Of all the papers that we examined in this section of the literature (perhaps 10 in total), only one, Stewart (2005), dwells on this issue at any length. As he notes, "Estimated coefficients in the standard parameterization of the Ordered Probit model cannot be interpreted directly and are only identified up to a scale normalization ... However, ratios of coefficients can be usefully interpreted." Strictly, this claim is correct when the partial effects in the true model obeys the "parallel regressions" feature and it is somewhat misleading as it only appies to a particular outcome – the partial effects change sign and magnitude as one moves through the set of outcomes. That is, when the partial effects are of the form $\partial\text{Prob}(y = j|\mathbf{x})/\partial x_k = K\boldsymbol{\beta}$ for some K that is independent of k. Stewart notes that this feature is useful for examining "indifference curves," that is, for examining what trades of two variables will leave the outcome (or, underlying preference) unchanged. [Boes and Winkelmann (2006a) pursue this same point at great length.] A second motivation for examining the ratios of coefficients is to see the ratios of specific partial effects, relative to a particular variable. He notes, in the Lewbel formulaton, one of the coefficients (that on the "special z") is normalized to 1. As such, each coefficient on another variable is interpretable as relative to this variable. Of course, the normalization could be on any other variable to secure identification of the model, but that would leave Stewart's observation intact. The ratios of coefficients on other variables to the z in question would survive renormalization of the model. However, even with all this in place, the analysis hangs on the assumption that the ratios of partial effects in the model equal the ratios of the parameters. In some of the model extensions we have examined, this is not the case.

The upshot of all this is that there is a loose end remaining to be tied up in the development of the semiparametric estimators. In the parametric formulations, the otherwise annoying scale difference between, say, probit and logit estimates is reconciled by the scaling of the model, itself. That reconciliation remains to be developed for the semiparametric approaches. This is needed in order to make the "robust" parameter estimates meaningful.

12.6 A Nonparametric Duration Model

Han and Hausman (1988, 1990) suggested a nonparametric approach to analysis of duration times that, after some manipulation, is treated in the framework of the ordered choice models considered here. [The model is also documented at length in Bhat (1996a,b) and Bhat and Pinjani (2008).]

Define T_i to be the time elapsed in an ongoing activity (use of a product under warranty, operation of a light bulb, life after transplant of a patient, economic life after layoff, duration of a trip, etc.) for individual i, measured on a continuous time scale. Observations in this setting will be a set of J+1 discrete intervals, $[0,t_1),[t_1,t_2),[t_2,t_3),\ldots,[t_J,+\infty)$, the time until transition (failure, death, exercise of the warranty, rehire, end of the trip, etc.) for individual i, and the interval, j, in which the transition takes place. We also assume that there is a set of measured covariates, $\mathbf{x}_i$, that remains fixed from the baseline until the ending period. Observations may be censored, which would be observationally equivalent to 'failing' in the rightmost interval or not failing at all during the observation period. Let y_i denote the observation on which interval contains the failure time, so $y_i = j$ if failure takes place between t_{j-1} and t_j. We develop a model for the determination of y_i.

The hazard function for the random variable T_i is

$$\lambda_i(\tau) = \lim_{\Delta \downarrow 0} \frac{\text{Prob}[\tau < T_i < \tau + \Delta \mid T_i > \tau]}{\Delta}. \tag{12.12}$$

The proportional hazards model specifies further that

$$\lambda_i(\tau) = \lambda_0(\tau) \exp(-\alpha - \boldsymbol{\beta}'\mathbf{x}_i), \tag{12.13}$$

where $\lambda_0(\tau)$ is the baseline hazard function. It follows that the log of the integrated hazard function may be expressed as

$$\log \int_0^{t_i} \lambda_0(\tau) d\tau = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \tag{12.14}$$

where ε_i has an extreme value distribution, $F(\varepsilon_i) = 1 - \exp(-\exp(\varepsilon_i))$. [There is a one to one correspondence between the log of the integrated hazard function and the density or the cdf, and the extreme value distribution corresponds to the preceding function. See Kalbfleisch and Prentice (2002).] For reasons that will emerge shortly, we have made the constant term in the hazard function explicit, whereas it is subsumed in $\boldsymbol{\beta}$ in Han and Hausman (1990). Further define

$$\mu_t = \log \int_0^t \lambda_0(\tau) d\tau, \; t = 1,...,T. \tag{12.15}$$

Then, the probability of failure in period j by individual i is

$$\text{Prob}[y_i = j \mid \mathbf{x}_i] = \int_{\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i}^{\mu_j - \boldsymbol{\beta}'\mathbf{x}_i} f(\varepsilon_i) d\varepsilon_i. \tag{12.16}$$

This is precisely the form of the probability for the ordered logit model. In more familiar terms,

$$\text{Prob}[y_i = j|\mathbf{x}_i] = F[\mu_j - \boldsymbol{\beta}'\mathbf{x}_i] - F[\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i], j = 0,1,\ldots,J. \tag{12.17}$$

The log likelihood function (as shown in Section 2.4) is

$$\log L = \sum_{i=1}^{n} \log \sum_{j=0}^{J} m_{ij} \left[\Lambda(\mu_j - \alpha - \boldsymbol{\beta}'\mathbf{x}_i) - \Lambda(\mu_{j-1} - \alpha - \boldsymbol{\beta}'\mathbf{x}_i) \right], \tag{12.18}$$

where $\Lambda(.)$ is the cdf of the logistic random variable and $m_{ij} = 1$ if y_i equals j and 0 otherwise. The integrated hazard functions and the slope parameters are estimable using maximum likelihood, simply by placing the observed interval observations into the ordered logit model framework. An application appears below.

12.6.1 Unobserved Heterogeneity

A considerable literature is devoted to accommodating unobserved individual heterogeneity in the model. [See Han and Hausman (1990) and Bhat (1996a) for extensive discussion.] Various parametric, semiparametric and nonparametric forms for the proportional hazards model can be conveniently assembled in the specification

$$\lambda_i(\tau|w_i) = \lambda_0(\tau) \exp(-\alpha - \boldsymbol{\beta}'\mathbf{x}_i + w_i). \tag{12.19}$$

The standard gamma model, in which $\exp(w_i)$ has a gamma density with mean 1 and variance σ^2, can be derived in closed form, as shown by Han and Hausman (1990). The log likelihood for the model with unobserved heterogeneity is obtained as

$$\begin{aligned}\log L &= \sum_{i=1}^{n} \log \int_{w_i} L(y_i : \boldsymbol{\beta}, \boldsymbol{\mu} \mid \mathbf{x}_i, w_i) f(w_i : \sigma^2) dw_i \\ &= \sum_{i=1}^{n} \log \int_{0}^{\infty} \left\{ \sum_{j=0}^{J} m_{ij} \left[\begin{matrix} \Lambda(\mu_j - \alpha - \boldsymbol{\beta}'\mathbf{x}_i - w_i) - \\ \Lambda(\mu_{j-1} - \alpha - \boldsymbol{\beta}'\mathbf{x}_i - w_i) \end{matrix} \right] \right\} \frac{\theta^{\theta} \exp(-\theta w_i) w_i^{\theta-1}}{\Gamma(\theta)} dw_i,\end{aligned} \tag{12.20}$$

where $\theta = 1/\sigma^2$. Let

$$I_i(t) = \exp(-\boldsymbol{\beta}'\mathbf{x}_i)\exp(\mu_t). \tag{12.21}$$

Then, the log likelihood function for the expanded model is (after considerable manipulation)

$$\log L = \sum_{i=1} \log\left[1 + \sigma^2 \Sigma_{t=0}^{y_i} I(t)\right]^{-1/\sigma^2} . \tag{12.22}$$

Aside from its mathematical convenience, however, it is difficult to motivate the log gamma model for latent heterogeneity. A model in which the heterogeneity is normally distributed, which would be more natural given that the heterogeneity is intended to capture latent characteristics of the individual, is now simple to devise, by using the random parameters specification presented in Section 8.1. The random parameters ordered logit model with only the constant term specified as a random parameter is consistent with the model for heterogeneity as an alternative to the log gamma model;

$$\lambda_i(\tau|w_i) = \lambda_0(\tau) \exp[-(\alpha + w_i) - \boldsymbol{\beta}'\mathbf{x}_i), w_i \sim N[0,\sigma^2]. \tag{12.23}$$

The model can be estimated by maximum simulated likelihood using the techniques developed in Section 8.1.2. The simulated log likelihood function for this model would be

$$\log L_S = \sum_{i=1}^{n} \log \frac{1}{R} \sum_{r=1}^{R} \left\{ \sum_{j=0}^{J} m_{ij} \left[\begin{matrix} \Lambda(\mu_j - \alpha - \boldsymbol{\beta}'\mathbf{x}_i - \sigma v_{i,r}) - \\ \Lambda(\mu_{j-1} - \alpha - \boldsymbol{\beta}'\mathbf{x}_i - \sigma v_{i,r}) \end{matrix} \right] \right\}, \tag{12.24}$$

where the simulation is over R replications, and $v_{i,r}$ is a draw from the standard normal population. By simply rearranging the function in terms of $\alpha_i = \alpha + \sigma v_i$, we obtain a restricted version of the random parameters model in Section 8.1 in which only the constant term is random. The example below extends the earlier application to allow for normally distributed individual heterogeneity.

Heckman and Singer (1984a,b) argued that a fully parametric model for the heterogeneity is likely to distort the estimated distributions. Their recommendation is consistent with a finite mixture (latent class) formulation, such as

$$\lambda_i(\tau)|q = \lambda_0(\tau) \exp[-\alpha_q - \boldsymbol{\beta}'\mathbf{x}_i], q = 1,\ldots,Q, \tag{12.25}$$

that is, a finite mixture, ordered logit model in which the threshold parameters and slopes are constant across classes, and the constant term defines the inter-class variation. The log likelihood function for this form of the model would be

$$\log L = \sum_{i=1}^{n} \sum_{q=1}^{Q} \pi_q \log\left\{\sum_{j=0}^{J} m_{ij}\left[\Lambda(\mu_j - \alpha_q - \boldsymbol{\beta}'\mathbf{x}_i) - \Lambda(\mu_{j-1} - \alpha_q - \boldsymbol{\beta}'\mathbf{x}_i)\right]\right\}, \quad (12.26)$$

where there are Q classes (support points in this instance), and π_q is the unconditional class probability (mass point in the heterogeneity distribution). As shown in Section 8.2, we can use a multinomial logit form to constrain the unconditional probabilities to sum to one;

$$\pi_q = \frac{\exp(\theta_q)}{\Sigma_{q=1}^{Q}\exp(\theta_q)}, \theta_Q = 0. \quad (12.27)$$

Bhat (1996a) suggests that the choice of Q be based on the smallest value of the Akaike Information Criterion,

$$\text{AIC}(Q) = -\log L - .5\, P \log N, \quad (12.28)$$

where P is the number of parameters estimated. (The parameter count would not include θ_q as this is a one to one function of the probabilities associated with the mass points, not a 'parameter.') Bhat's 1996a application studied the duration of shopping trips by households in a survey conducted in April of 1991 by the Central Transportation Planning Staff (CTPS) in the Boston Metropolitan region.

12.6.2 Application

To illustrate the use of the ordered logit model to study duration data, we will examine Kennan's (1985) data on the duration of 62 strikes in the U.S. from 1968 to 1976. The data consist of the 62 durations and, for each year, a measure of "unanticipated" aggregate output. [These data are Table F25.3 in Greene (2008, Appendix F).] The data on strike durations and the index of unanticipated output are shown in Figure 12.3. The durations were arbitrarily grouped (purely for this example) into the nine intervals shown in Table 12.1.

The estimated ordered logit model is shown in Table 11.2. The results suggest that unanticipated production is a somewhat significant influence on the hazard rate for strike duration. To use these results to examine the hazard rates, we compute the following estimates of the hazard functions:

$$\hat{\lambda}(t_j) = \frac{\left[\Lambda\left(\hat{\mu}_j - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}\right) - \Lambda\left(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}\right)\right]}{\left[1 - \Lambda\left(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}\right)\right]\left(\hat{\mu}_j - \hat{\mu}_{j-1}\right)}. \quad (12.29)$$

The results are shown in Figure 12.4. The jagged imprecision of the estimated function is likely due to the small sample (cell) sizes and the small number of intervals considered.

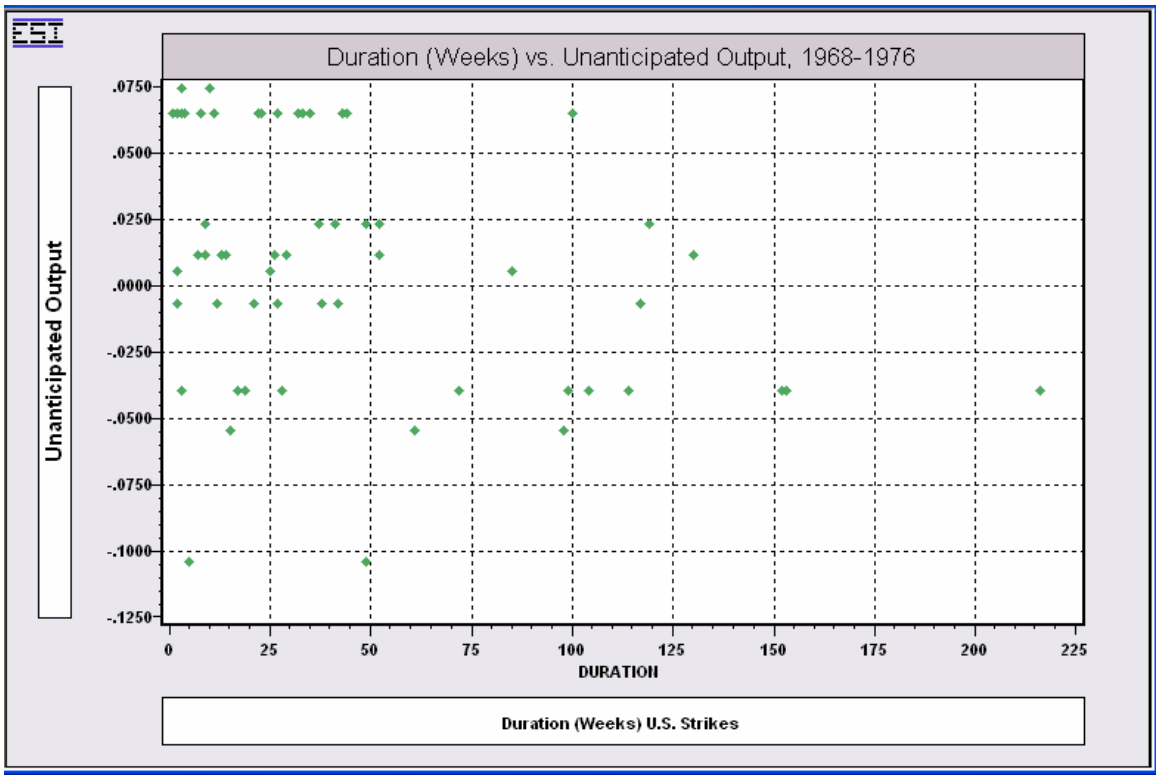


Figure 12.3 Strike Duration Data

Table 12.1. Grouping of Strike Durations

j	Duration	Frequency Count	Frequency Percent	Cumulative < = Count	Cumulative < = Percent	Cumulative > = Count	Cumulative > = Percent
0	0 to 4 weeks	10	16.1290	10	16.1290	62	100.0000
1	5 to 10 weeks	6	9.6774	16	25.8065	52	83.8710
2	11 to 13 weeks	4	6.4516	20	32.2581	46	74.1935
3	14 to 17 weeks	3	4.8387	23	37.0968	42	67.7419
4	18 to 23 weeks	5	8.0645	28	45.1613	39	62.9032
5	24 to 28 weeks	5	8.0645	33	53.2258	34	54.8387
6	29 to 40 weeks	6	9.6774	39	62.9032	29	46.7742
7	41 to 60 weeks	9	14.5161	48	77.4194	23	37.0968
8	61 to ∞ weeks	14	22.5806	62	100.0000	14	22.5806

Since we have the actual realizations of T_i in hand, we can estimate the logistic hazard model that is being approximated with the ordered logit model. The density and hazard functions for the log-logistic model are

$$f(T_i) = \frac{(\delta_i p)(\delta_i T_i)^{p-1}}{[1+(\lambda_i T_i)^p]^2},$$

$$\lambda(T_i) = \frac{f(T_i)}{[1+(\delta_i T_i)^p]}, \tag{12.30}$$

where $\delta_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$, p is a scale parameter and the denominator of the hazard function is the survival function, Prob[$t_i \geq T_i$]. The estimated parameters are shown in Table 12.2. The resemblance to the ordered logit estimates is to be expected, since the latter are a (rough) approximation The hazard function for the parametric model is shown in Figure 12.5.

Finally, the estimates of an ordered logit model with (log)normal heterogeneity accounted for in the hazard function are also shown in Table 12.2. The estimates are obtained by maximum simulated likelihood, as discussed earlier. The results show little evidence of heterogeneity. Of course this is to be expected, since the boundaries of the time intervals are set arbitrarily, and these are aggregate data in any event. (The log likelihood functions for the two ordered choice models are almost identical.)

Table 12.2 Estimated Logistic Duration Models for Strike Duration

Variable	Coeff.	Standard Error	b/St.Er	Prob.	Coeff.	Standard Error	b/St.Er.	Prob.
	Ordered logit model				Normal heterogeneity model			
Constant	1.6088	.2416	6.658	.0000	1.6087	.3467	4.641	.0000
PROD	7.9272	4.7790	1.659	.0971	7.9287	5.8621	1.353	.1762
	Threshold parameters for index							
Mu(1)	.6115	.2006	3.049	.0023	.6115	.2419	2.533	.0113
Mu(2)	.9491	.2175	4.363	.0000	.9491	.2832	3.351	.0008
Mu(3)	1.1886	.2253	5.276	.0000	1.1886	.3085	3.853	.0001
Mu(4)	1.5692	.2333	6.727	.0000	1.5692	.3295	4.763	.0000
Mu(5)	1.9257	.2413	7.981	.0000	1.9257	.3430	5.615	.0000
Mu(6)	2.3413	.2553	9.171	.0000	2.3413	.3597	6.509	.0000
Mu(7)	3.0358	.3041	9.984	.0000	3.0358	.4051	7.494	.0000
Sigma					.0042	.2274	.019	.9852
	Estimated log-logistic parametric hazard model							
Constant	3.0256	.1594	18.977	.0000	Number of Observations = 62			
PROD	6.4482	4.2292	1.525	.1273	Mean of PROD = 0.0110231			
p	1.5233	.1919	7.934	.0000				

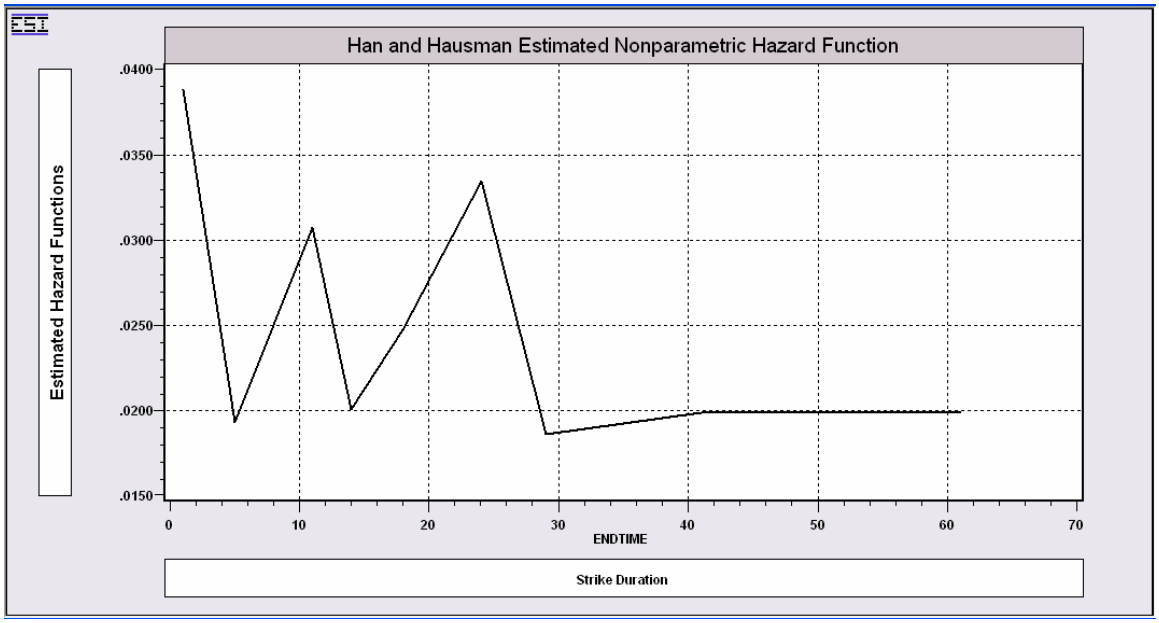


Figure 12.4 Estimated Nonparametric Hazard Functions

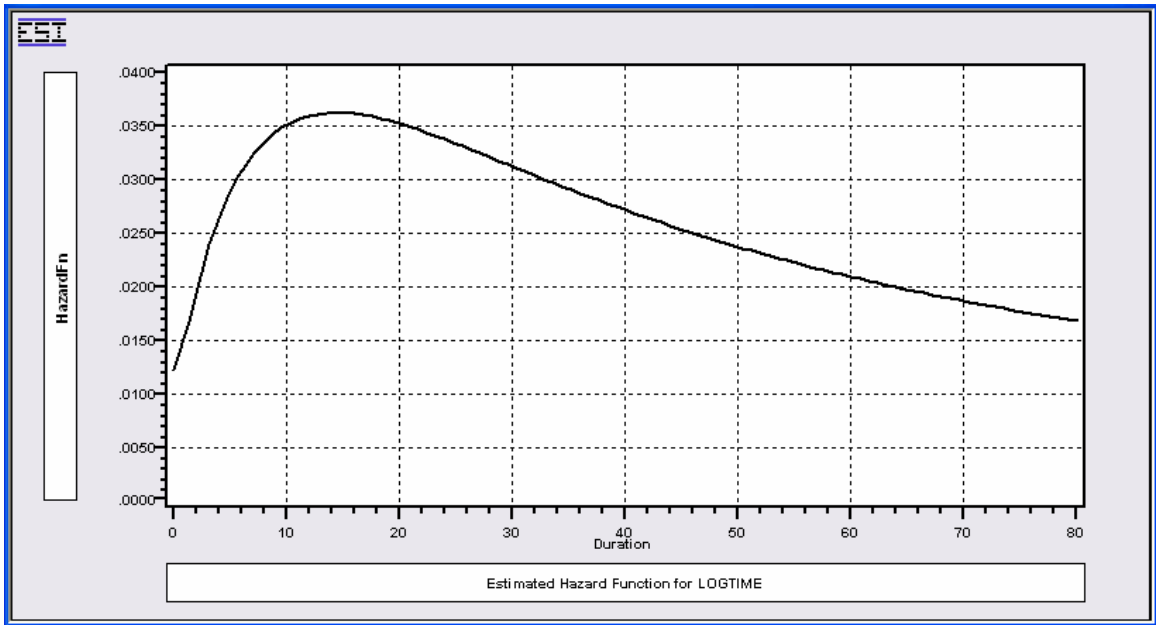


Figure 12.5 Estimated Hazard Function from Loglogistic Parametric Model

References

Abramovitz, M. and I. Stegun, 1971. *Handbook of Mathematical Functions*, New York, Dover Press.

Abrevaya, J., 1997. "The Equivalence of Two Estimators of the Fixed Effects Logit Model," *Economics Letters*, 55, pp. 41-43.

Acharya, S., 1988. "A Generalized Econometric Model and Tests of a Signalling Hypothesis with Two Discrete Signals," *Journal of Finance*, 43, pp. 413-429.

Adams, J., 2006. "Learning, Internal Research and Spillovers," *Economics of Innovation and New Technology*, 15, pp. 5-36.

Agresti, A., 1984. *Analysis of Ordinal Categorical Data*, New York, Wiley.

Agresti, A., 1990. *Categorical Data Analysis*, New York, John Wiley and Sons.

Agresti, A., 1999. "Modelling Ordered Categorical Data: Recent Advances and Future Challenges," *Statistics in Medicine*, 18, pp. 2191-2207.

Aguemang-Duah, K. and F. Hall (1997) Spatial Transferability of an Ordered Response Model of Trip Generation, Transport Research – Series A, 31, 5, 389-402

Agostino, A., C. Bhat and E. Pas (1996) A Random Effects Multinomial Probit Model of Car Ownership Choice, Proceedings of the Third Workshop on Bayesian Statistics in Science and Technology, Cambridge University Press.

Aitchison, J. and J. Bennett, 1970. "Polychotomous Quantal Response by Maximum Indicant," *Biometrika*, 57, pp. 253-262.

Aitcheson, J. and S. and Silvey, 1957. "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika*, 44, pp. 131-140.

Albert, J. and S. Chib, 1993. "Bayesian Analysis of Binary and Polytomous Response Data," *Journal of the American Statistical Association*, 88, pp. 669-679.

Aldrich, J., and F. Nelson., 1984. *Linear Probability, Logit, and Probit Models*. Beverly Hills, Sage Publications.

Allison, P., 1999. "Comparing Logit and Probit Coefficients Across Groups," *Sociological Methods and Research*, 28,, pp. 186-208.

Amel, D. and J. Liang, 1994. "A Dynamic Model of Entry and Performance in the U.S. Banking Industry," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, 210.

Amel, D. and J. Liang, 1997. "Determinants of Entry and Profits in Local Banking Markets," *Review of Industrial Organization*, 12, pp. 59-78.

Amemiya, T., 1975. "Qualitative Response Models," *Annals of Economic and Social Measurrement*, 4, pp. 363-372.

Amemiya, T., 1980. "The n^{-2} – order Mean Squared Errors of the Maximum Likelihood and the Minimum Logit Chi Squared Estimator," *Annals of Statistics*, 8, pp. 488-505.

Amemiya, T. 1981. "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19, 4, pp. 481-536.

Amemiya, T., 1985a. "Tobit Modeling: A Survey," *Journal of Econometrics*, 24, 1/2, pp. 3-61.

Amemiya, T., 1985b. *Advanced Econometrics*, Cambridge, Harvard University Press

Ananth, C. and D. Kleinbaum, 1997. "Regression Models for Ordinal Responses: A Review of Methods and Applications," *International Journal of Epidemiology*, 26, pp. 1232-1333.

Andersen, D., 1970. "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the royal Statistical Society, Series B*, 32, pp. 283-301.

Anderson, J., 1972. "Separate Sample Logistic Discrimination," *Biometrika*, 59, pp. 19-35.

Anderson, J., 1984. "Regression and Ordered Categorical Variables," *Journal of the Royal Statistical Society, Series B (Methodological)*, 46, pp. 1-30.

Anderson, J. and P. Philips, 1981. "Regression, Discrimination and Measurement Models for Ordered Categorical Variables," *Applied Statistics*, 30, pp. 22-31.

Ando, T., 2006. "Bayesian credit rating analysis based on ordered probit regression model with functional predictor," *Proceeding of The Third IASTED International Conference on Financial Engineering and Applications,* 69-76.

Andrews, D. and W. Ploberger ,1994. "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 62, pp. 1383-1414.

Andrich, D., 1979, "A Model for Contingency Tables Having an Ordered Response Classification, *Biometrics*, 35, 403-415.

Angelini, V., Cavapozzi, D., Corazzini, L. and O. Paccagnella,. 2008.." Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Speci.c Scale Biases?. Manuscript. University of Padua.

Angrist, J., 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 29, 1, pp. 2-15.

Ansari, A., S. Essegaier and R. Kohli, 2000. "Internet Recommendation Systems,"*Journal of Marketing Research*, 37, 3, 363-375.

Arellano, M., 2001. "Panel Data: Some Recent Developments," in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, Volume 5, North Holland, Amsterdam.

Armstrong. D. and J. McVicar, 2000. "Value Added in Further Education and Vocational Training in Northern Ireland," *Applied Economics*, 32, pp. 1727-1736.

Avery, R., L. Hansen, and J. Hotz, 1983. "Multiperiod Probit Models and Orthogonality Condition Estimation." *International Economic Review,* 24, pp. 21–35.

Bago d.Uva, T., E. Doorslear.v., M. Lindeboom, M. and O.Donnell, 2008. "Does Reporting Heterogeneity Bias the Measurement of Health Disparities?" *Health Economics*, 17, 351-375.

Barnhart, H. and A. Sampson, 1994. "Overview of Multinomial Models for Ordered Data," *Communications in Statistics – A. Theory and Methods*, 23, pp. 3395-3416.

Baltagi, B., 2005. *Econometric Analysis of Panel Data, 3rd ed.,* New York, John Wiley and Sons.

Baltagi, B., 2007. *Econometric Analysis of Panel Data, 4th ed.,* New York, John Wiley and Sons.

Basu, D. and R. de Jong, 2006. "Dynamic Multinomial Ordered Choice With An Application to the Estimation of Monetary Policy Rules," Department of Economics, Ohio State University, Manuscript.

Becker, W. and P. Kennedy, 1992. "A Graphical Exposition of the Ordered Probit Model," *Econometric Theory*, 8, pp. 127-131.

Bedi, A. and I. Tunali, 2004. "Testing for Market Imperfections: Participation in Land and Labor Contracts in Turkish Agriculture," Working Paper, Institute of Social Studies, The Hague.

Bellemare, C., B. Melenbert and A. van Soest, 2002. "Semi-parametric Models for Satisfaction with Income," *Portuguese Economic Journal*, 1, pp. 181-203.

Ben-Akiva, M., and S. Lerman, 1985. *Discrete Choice Analysis*. London: MIT Press.

Bennett, J. and S. Lanning, 2007. "The Netflix Prize," *Proceedings of the KDD Cup and Workshop*.

Bera, A., C. Jarque and L. Lee, 1984. "Model Specification Tests: A Simultaneous Approach," *Journal of Econometrics*, 20, pp. 59-82.

Berndt, E., B. Hall, R. Hall and J. Hausman, 1974. "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3/4, pp. 653-665.

Berkson, J., 1944. "Application of the Logistic Function to Bioassay," *Journal of the American Statistical Association*, 39, pp. 357-365.

Berkson, J., 1951. "Why I Prefer Logits to Probits," *Biometrics*, 7, 4, pp. 327-339.

Berkson, J., 1953. "A Statistically Precise and Relatively Simple Method of Estimating the Bioassay with Quantal Response, Based on the Logistic Function," *Journal of the American Statistical Association*, 48, pp. 565-599.

Berkson, J., 1955a. "Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function," *Journal of the American Statistical Assocation*," 50, pp. 130-162.

Berkson, J., 1955b. "Estimate of the Integrated Normal Curve by Minimum Normit Chi-Square with Particular Reference to Bioassay," *Journal of the American Statistical Assocation*," 50, pp. 529-550.

Berkson, J., 1957. "Tables for Use In Estimating the Normal Distribution Function by Normit Analysis," *Biometrika*, 44, pp. 411-435.

Berkson, J., 1980. "Minimum Chi-Square, Not Maximum Likelihood," *Annals of Statistics*, 8, pp. 457-487.

Bertschek, I. and M. Lechner, 1998. "Convenient Estimators for the Panel Probit Model," *Journal of Econometrics*, 87, 2, pp. 329-372

Bhat, C., 1994. "Imputing a Continuous Income Variable from Grouped and Missing Income Observartions," *Economics Letters*, 46, 4, pp. 311-320.

Bhat, C., 1996a. "A Hazard-Based Duration Model of Shopping Activity with Nonparametric Baseline Specification and Nonparametric Control for Unobserved Heterogeneity," *Transportation Research Part B*, 30(3), 189-207.

Bhat, C., 1996b. "A Generalized Multiple Durations Proportional Hazard Model with an Application to Activity Behavior During the Work-to-Home Commute," *Transportation Research Part B*, 30, 465-480.

Bhat, C., 1997. "Work Travel Mode Choice and Number of Nonwork Commute Stops," *Transportation Research Part B*, 31(1), 41-54.

Bhat, C., 1999. "An Analysis of Evening Commute Stop-Making Behavior Using Repeated Choice Observations from a Multi-Day Survey," *Transportation Research Part B*, 33(7), 495-510.

Bhat, C., 2001. "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," *Transportation Research Part B*, 35(7), 677-693.

Bhat, C., 2003. "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences," *Transportation Research Part B*, 37(9), 837-855.

Bhat, C., J. Carini and R. Misra (1999) "Modeling the Generation and Organization of Household Activity Stops. *Transportation Research Record*, 1676, 153-161.

Bhat, C. and F. Koppelman (1993) An Endogenous Switching Simultaneous Equation System of Employment, Income and Car Ownership, *Transportation Research A*, 27, 447-459.

Bhat, C. and A. Pinjari, 2008. "Duration Modeling," *Handbook of Transport Modelling, 2nd edition*, Chapter 6, pp. 105-132, edited by D.A. Hensher and K.J. Button, Elsevier Science.

Bhat, C. and V. Pulugurta, 1998. "A Comparison of Two Alternative Behavioral Mechanisms for Car Ownership Decisions," *Transportation Research Part B*, 32(1), 61-75.

Bhat, C. and S. Singh, 2000. "A Comprehensive Daily Activity-Travel Generation Model System for Workers," *Transportation Research Part A*, 34(1), 1-22.

Bhat, C. and S. Srinivasan, 2005. "A Multidimensional Mixed Ordered-Response Model for Analyzing Weekend Activity Participation," *Transportation Research Part B*, 39(3), 255-278.

Bhat, C. and H. Zhao, 2002. "The Spatial Analysis of Activity Stop Generation," *Transportation Research Part B*, 36(6), 557-575.

Biswas and Das, 2002. "A Bayesian Analysis of Bivariate Ordinal Data: Wisconsin Epidemiologic Study of Diabetic Retinopathy Revisited, *Statistics in Medicine*, 21, 4, pp. 549-559

Bliss, C., 1934a. "The Method of Probits," *Science*, 79, 2037, pp. 38-39.

Bliss, C., 1934b. "The Method of Probits: A Correction," *Science*, 79, 2053, pp. 409-410.

Blundell, R., F. Laisney, and M. Lechner, 1993. "Alternative Interpretations of Hours Information in an Econometric Model of Labour Supply." *Empirical Economics,* 18, pp. 393–415.

Blundell, R. and J. Powell, 2004. "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 7, pp. 655-679.

Boes, S., 2007. "Nonparametric Analysis of Treatment Effects in Ordered Response Models," University of Surich, Socioeconomic Institute, Working Paper 0709.

Boes, S. and R. Winkelmann, 2004. "Income and Happiness: New Results from Generalized Threshold and Sequential Models," IZA Discussion Paper No. 1175, SOI Working Paper 0407, IZA

Boes, S. and R. Winkelmann, 2006a. "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 1, pp. 165-180.

Boes, S. and R. Winkelmann, 2006b. "The Effect of Income on Positive and Negative Subjective Well-Being," University of Zurich, Socioeconomic Institute, Manuscript, IZA Discussion Paper Number 1175.

Boyes, W., D. Hoffman and S. Low, 1989. "An Econometric Analysis of the Bank Credit Scoring Problem," *Journal of Econometrics*, 40, pp. 3-14.

Brant, R., 1990. "Assessing Proportionality in the Proportional Odds Model for Ordered Logistic Regression," *Biometrics*, 46, pp. 1171-1178.

Bresnahan, T. F., 1987. "Competition and Collusion in the American Automobile Industry: The 1955 Price War," *Journal of Industrial Economics*, 35, pp. 457–482.

Breusch, T. and A. Pagan, 1979. "A Simple Test for Heteroscedasticity and Random Parameter Variation," *Econometrica*, 47, pp. 1287-1294.

Brewer, C., C. Kovner, W. Greene, Y. Cheng, 2008. **"**Predictors of RNs' Intent to Work and Work Decisions One Year Later in a U.S. National Sample," *The International Journal of Nursing Studies*, forthcoming.

Bricka, S., and C. Bhat, 2006. "A Comparative Analysis of GPS-Based and Travel Survey-based Data,". *Transportation Research Record*, 1972, 9-20.

Britt, C., 2000. "Comment on Paternoster and Brame," *Criminology*, 38, 3, pp. 965–970**.**

Brooks, R., M. Harris and C. Spencer, 2007. "A Inflated Ordered Probit Model of Monetary Policy: Evidence from MPC Voting Data," Department of Econometrics and Business Statistics, Monash University, Manuscript.

Buckle, R. and J. Carlson 2000, 2000. "Menu Costs, Firm Size and Price Rigidity," *Economics Letters*, 66, pp. 59-63.

Buliung, R., 2005. "Activity/Travel Behaviour Research: Approaches and Findings with Identification of Researching Themes and Emerging Methods," Center for Spatial Analysis, McMaster Univ, Working paper 008. (http://www.science.mcmaster.ca/cspa/papers/CSpA%20WP%20008.pdf).

Bunch, D. and R. Kitamura (1990) Multinomial Probit Estimation Revisited: Testing Estimable Model Specifications, Maximum Likelihood Algorithms and Probit Integral Approximations for Car Ownership, Institute for Transortation Studies Technical Report, University of California, Davis.

Burnett, N. "Gender Economics Courses in Liberal Arts Colleges." *Journal of Economic Education,* 28, 4, 1997, pp. 369–377.

Butler, J., T. Finegan and J. Siegfried, 1994. "Does More Calculus Improve Student Learning in Intermediate Micro and Macro Economic Theory?" *American Economic Review*, 84, 2, pp. 206-210.

Butler, J., T. Finegan and J. Siegfried, 1998. "Does More Calculus Improve Student Learning in Intermediate Micro- and Macroeconomic Theory?" *Journal of Applied Econometrics*, 13, pp. 185-202.

Butler, J. and P. Chatterjee, 1995. "Pet Econometrics: Ownership of Cats and Dogs," Department of Economics, Vanderbilt University, Working Paper Number 95-WP1.

Butler, J. and P. Chatterjee, 1997. "Tests of the Specification of Univariate and Bivariate Ordered Probit," *Review of Economics and Statistics*, 79, pp. 343-347.

Butler, J. and R. Moffitt, 1982. "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model," *Econometica*, 50, pp. 761-764.

Calhoun, C. A. 1986. "BIVOPROB: Maximum Likelihood Program for Bivariate Ordered-Probit Regression." Washington, D.C.: The Urban Institute.

Calhoun, C., 1989. "Estimating the Distribution of Desired Family Size and Excess Fertility," *Journal of Human Resources*, 24, 4, pp. 709-24.

Calhoun, C., 1991. "Desired and Excess Fertility in Europe and the United States: Indirect Estimates from World Fertility Survey Data," *European Journal of Population*, 7, pp. 29-57.

Calhoun, C., 1994. "The Impact of Children on the Labor Supply of Married Women: Comparative Estimates from European and U.S. Data," *European Journal of Population*, 10, pp. 293-318.

Calhoun, C., 1995. "BIVOPROB: Computer Program for Maximum Likelihood Estimation of Bivariate Ordered Probit Models for Censored Data: Version 11.92," *Economic Journal*, 105, pp. 786-787.

Cameron, A. and P. Trivedi (1998) Regression Analysis of Count Data, New York, Cambridge University Press.

Cameron, A. and P. Trivedi (2005) Microeconometrics: Methods and Applications, Cambridge, Cambridge University Press.

Cameron, S. and J. Heckman, 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, pp. 262–333.

Carneiro, P., K. Hansen and J. Heckman, 2001. "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies," *Swedish Economic Policy Review*, 8, pp. 273-301.

Carneiro, P., K. Hansen and J. Heckman, 2003. "Estimating Distributions of Treatment Effects with an Application to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, pp. 361-422.

Carro, J., 2007. "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," *Journal of Econometrics*, 140, pp. 503-528.

Caudill, S., 1988. "An Advantage of the Linear Probability Model Over Probit or Logit." *Oxford Bulletin of Economics and Statistics,* 50, pp. 425–427.

Cecchetti, S., 2006. "The Frequency of Price Adjustment: A Study of the Newsstand Prices of Magazines," *Journal of Econometrics*, 31, 3, pp. 255-274.

Chamberlain, G., 1980. "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, pp. 225-238.

Chen, S. and S. Khan , 2003. "Rates of Convergence for Estimating Regression Coefficientsd in Heteroskedastic Discrete Response Models," *Journal of Econometrics*, 117, pp. 245-278.

Chesher, A. and M. Irish, 1987. "Residual Analysis in the Grouped Data and Censored Normal Linear Model," *Journal of Econometrics*, 34, pp. 33-62.

Cheung, S., 1996. "Provincial Credit Rating in Canada: An Ordered Probit Analysis," Bank of Canada, Working Paper 96-6. (http://www.bankofcanada.ca/en/res/wp/1996/wp96-6.pdf)

Chow, G., 1960. "Testss of Equality Between Sets of Coefficients in two Linear Regressions," *Econometica*, 28, pp. 591-605.

Christensen, K., H. Kohler, O, Basso, Olga, J. Olsen, J. Vaupel and J. Rodgers, 2003. "The Correlation of Fecundability Among Twins: Evidence of a Genetic Effect on Fertility?" *Epidemiology*, 14, 1, pp. 60-64.

Christofides, L., T. Stengos, and R Swidinsky, 1997. "On the Calculation of Marginal Effects in the Bivariate Probit Model." *Economics Letters,* 54, 3, pp. 203–208.

Christofides, L., T. Hardin, and R. Stengos, 2000. "On the Calculation of Marginal Effects in the Bivariate Probit Model: Corrigendum." *Economics Letters,* 68, pp. 339–340.

Clark, A., Y. Georgellis and P. Sanfey, 2001. "Scarring: The Psychological Impact of Past Unemployment," *Economica*, 68, pp. 221-241. et al. 2001

Clogg, C. and E. Shihadeh, 1994. *Statistical Models for Ordered Variables*, Thousand Oaks, CA, Sage Publications.

Contoyannis, A., A. Jones and N. Rice, 2004. "The Dynamics of Health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19, 4, pp. 473-503.

Coppejans, M., 2007. "On Efficient Estimation of the Ordered Response Model," *Journal of Econometrics*, 137, pp. 577-614.

Cox, C., 1995. "Location-Scale Cumulative Odds Models for Ordered Data: A Generalized Nonlinear Model Approach," *Statistics in Medicine*, 14, pp. 1191-1203.

Cox, D., 1970. *Analysis of Binary Data*, Methuen, London.

Cragg, J. (1971) Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods, Econometrica, 39, 829-844.

Cragg, J., and R. Uhler, 1970. "The Demand for Automobiles." *Canadian Journal of Economics,* 3, pp. 386–406.

Cramer, J., 1999. "Predictive Performance of the Binary Logit Model in Unbalanced Samples." *Journal of the Royal Statistical Society, Series D (The Statistician)* 48, pp. 85–94.

Crawford, D., R. Pollak and F. Vella, 1988. "Simple Inference in Multinomial and Ordered Logit," *Econometric Reviews*, 17, pp. 289–299.

Crouchley, R., 1995. "A Random Effects model for Ordered Categorical Data," *Journal of the American Statistical Association*, 90, pp. 489-498.

Crouchley, B., 2005. "E-Science, The GRID and Statistical Modelling in Social Research," (http://www.ccsr.ac.uk/methods/festival/programme/gss/crouchley.ppt).

Cunha, F., J. Heckman and S. Navarro, 2007. "The Identification & Economic Content of Ordered Choice Models with Stochastic Thresholds," University College Dublin, Gery Institute, Discussion Paper WP/26/2007.

Czado, C., A. Heyn and G. Müller, 2005. "Modeling Migraine Severity with Autoregressive Ordered Probit Models," Technische Universitat, Munchen, Working paper number 463.

D'Addio, A., T. Eriksson and P. Frijters, 2007. "An Analysis of the Determinants of Job Satisfaction When Individuals' Baseline Satisfaction Levels May Differ," *Applied Economics*, 39, 19, pp. 2413-2423.

Dardanomi, V. and A. Forcina, 2004. "Multivariate Ordered Logit Regressions," University of Palermo, Manuscript. (http://www.cide.info/conf_old/papers/11128.pdf)

Das, M. and A. van Soest, 2000. "A Panel Data Model for Subjective Information on Household Income Growth," *Journal of Economic Behavior and Organization*, 15, pp. 401-416.

Davidson, R. and J. MacKinnon, 1983. "Small Sample Properties of Alternative Forms of the Lagrange Multiplier Test," *Economics Letters*, 12, pp. 269-275.

Davidson, R. and J. MacKinnon, 1984. "Model Specification Tests Based on Artificial Linear Regressions," *International Economic Review*, 25, pp. 485-502.

Davidson, R. and J. MacKinnon, 1993. *Estimation and Inference in Econometrics*, Oxford, Oxford University Press.

Daykin, A. and P. Moffatt , 2002. "Analyzing Ordered Responses: A Review of the Ordered Probit Model," *Understanding Statistics*, I, 3, pp. 157-166.

DeMaris, A., 2004. *Regression with Social Data: Modeling Continuous and Limited Response Variables*, Hoboken, New Jersey, John Wiley and Sons.

Dempster, A., N. Laird and D. Rubin 1977. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, pp. 1-38.

Diggle, R., P. Liang and S. Zeger, 1994. *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Drezner, Z., 1978. "Computation of the Bivariate Normal Integral," *Mathematics of Computation*, 32, 141, pp. 277-279.

Drezner, Z., 1994. "Computation of the Trivariate Normal Integral," *Mathematics of Computation*, 62, 205, pp. 289-294.

Dueker. M., S. Spuirr, A. Jacox and D. Kalist, 2005 "The Practice Boundaries of Advanced Practice Nurses: An Economic and Legal Analysis," Federal Reserve Bank of St. Louis**,** Working Paper 2005-071A

Dupor, B., T. Mirzoev, T. Conley, T., 2004. "Does the Federal Reserve Do What It Says It Expects to Do?" Working Paper, Department of Economics, Ohio State University, Manuscript.

Econometric Software, 2007. *NLOGIT: Version 4.0*, Plainview, New York.

Efron, B., 1978. "Regression and ANOVA with Zero-One Data: Measures of Residual Variation." *Journal of the American Statistical Association,* 73, pp. 113–212.

Eichengreen, B., M. Watson and R. Grossman, 1985. "Bank Rate Policy Under the Interwar Gold Standard: A Dynamic Probit Approach," *Economic Journal*, 95, pp. 725-745.

Ekholm, A. and J. Palmgren, 1989. "Regression Models for an Ordinal Response Are Best Handled As Nonlinear Models," *GLIM Newsletter*, 18, pp. 31-35.

Eluru, N., C. Bhat and D. Hensher, 2008. "A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Levels in Traffic Crashes," *Accident Analysis and Prevention*, 40, 3, pp. 1033-1054..

Everitt, B., 1988. A Finite Mixture Model for the Clustering of Mixed-Mode Data," *Statistics and Probability Letters*, 6, pp. 305-309.

EViews, 2008. *Eviews Version 6.0*, QMS, Irvine, CA.

Fahrmeir, L. and G. Tutz, 1994. *Multivariate Statistical Modeling Based on Generalized Linear Models*, Berlin, Springer Verlag.

Farewell, V., 1982. "A Note on Regression Analysis of Ordinal Data with Variability of Classification," *Biometrika*, 69, pp. 533-538.

Feinberg, S., 1980. *The Analysis of Cross-Classified Categorical Data*, Cambridge, MIT Press.

Ferrer-i-Carbonell, A. and P. Frijters, 2004. "How Important Is Methodology for the Estimates of the Determinants of Happiness," *Economic Journal*, 114, pp. 641-659.

Fernandez, A., and J. Rodriguez-Poo, 1997. "Estimation and Testing in Female Labor Participation Models: Parametric and Semiparametric Models." *Econometric Reviews,* 16, pp. 229–248.

Fernandez-Val I., 2008. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *Journal of Econometrics*, Forthcoming

Fernandez, A., and J. Rodriguez-Poo, 1997. "Estimation and Testing in Female Labor Participation Models: Parametric and Semiparametric Models," *Econometric Reviews*, 16, pp. 229-248.

Fernández-Val, I. and F. Vella, 2007. "Bias Corrections for Two-Step Fixed Effects Panel Data Estimators," IZA Working Papers Number 2690.

Filer, R. and M. Honig, 2005. "Endogenous Pensions and Retirement Behavior, Department of Economics, Hunter College, Manuscript.

Finney, D. 1944a. "The Application of the Probit Method to Toxicity Test Data Adjusted for Mortality in the Control", *Annals of Applied Biology*, 31, pp.68-74.

Finney, D., 1944b. "The Application of Probit Analysis to the Results of Mental Tests", *Psychometrika*, 9, pp. 31-39.

Finney, D. 1947, "The Principles of Biological Assays", *Journal of the Royal Statistical Association B*, 9, pp. 46-91.

Finney, D., 1947. *Probit analysis: A Statistical Treatment of the Sigmoid Response Curve,* Cambridge: Cambridge University Press.

Finney, D., 1952. *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve, 2nd Edition,* Cambridge: Cambridge University Press.

Finney, D., 1971. *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve, 3rd Edition,* Cambridge: Cambridge University Press.

Formisiano, J., K. Still, W. Alexander and M. Lippmann, 2001. "Applicatioin of Statistical Models for Secondary Data Usage of the U.S. Navy's Occupational Exposure Database (NOED)," *Applied Occupational and Environmental Hygiene*, 16, pp. 201-209.

Frazis, H., 1993. "Selection Bias and the Degree Effect," *Journal of Human Resources*, 28, pp. 538-554.

Freedman, D., 2006. "On the So-Called 'Huber Sandwich Estimator' and Robust Standard Errors," *The American Statistician*, 60, 4, pp. 299-302.

Frijters, P., J. Haisken-DeNew and M. Shields, 2004. "The Value of Reunification in Germany: An Analysis of Changes in Life Satisfaction," *Journal of Human Resources*, 39, 3, pp. 649-674.

Fu, V., 1998. "Estimating Generalized Ordered Logit Models," *Stata Technical Bulletin*, 44, pp. 27-30.

Fu, A., M, Gordon, G. Liu B. Dale and R. Christensen, 2004. "Inappropriate Medication Use and Health Outcomes in the Elderly" *Journal of the American Geriatrics Society*, 52, 11, pp. 1934–1939.

Gallant, R. and D. Nycjka, 1987. "Semi-nonparametric Maximum Likelihood Estimation," Econometrica, 55, pp. 363-390.

Garen, J., 1984. "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica*, 52, 5, pp. 1199-1218.

Genius, M., C. Pantzios, V. Tzouvelakis, 2005. "Information Acquisition and Adoption of Organic Farming Practices: Evidence from Farm Operations in Crete, Greece" Department of Economics, University of Crete, Manuscript.

Gaddum, J., 1933. "Reports on Biological Standards, III. Methods of Biological Assay Depending on a Quantal Response" Special Report Series 183. Medical Research Council, HM Statistical Office, London.

Gallant, R. and D. Nychka, 1987. "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, pp. 363-390.

Genberg, H. and S. Gerlach, 2004. "Estimating Central Bank Reaction Functions with Ordered Probit: A Note" Graduate Institute of International Studies, Geneva, Manuscript.

Genz, A., 2008. "Numerical Computation of Bivariate and Trivariate Normal Probabilities," Department of Mathematics Washington State University http://www.sci.wsu.edu/math/faculty/genz/papers/bvnn/bvnn.html)

Gerfin, M., 1996. "Parametric and Semi-Parametric Estimation of the Binary Response Model." *Journal of Applied Econometrics,* 11, pp. 321–340.

Geweke, J., 1991. Efficient Simulation From the Multivariate Normal and Student t-Distributions Subject to Linear Constraints," in *Computer Sciences and Statistics Proceedings of the 23d Symposium on the Interface*, pp. 571-578.

Girard, P. and E. Parent, 2001. "Bayesian Analysis of Autocorrelated Ordered Categorical Data for Industrial Quality Monitoring," *Technometrics*, 43, 2, pp. 180-191.

Glewwe, P., 1997. "A Test of the Normality Assumption in the Ordered Probit Model," *Econometric Reviews*, 16, 1, pp. 1-19.

Glewwe, P. and H. Jacoby, 1994. "Student Achievement and Schooling Chioce in Low-Income Countries: Evidence from Ghana," *Journal of Human Resources*, 29, 3, pp. 843-864.

Glewwe, P. and H. Jacoby, 1995. "An Economic Analysis of Delayed Primary School Enrollment in a Low Income Country: The Role of Early Childhood Malnutrition," *Review of Economics and Statistics*, 77, 1, pp. 156-169.

Godfrey, L., 1988. *Misspecification Tests in Econometrics*, Cambridge, Cambridge University Press.

Golob, T. (1990) The Dynamics of Household Travel time Expenditures and Car Ownership Decisions, Transportation Research A, 24, 443-465.

Golob, T. and L. van Wissen (1998) A Joint Household Travel Distance Generation and Car Ownership Model, Working Paper WP-88-15, Institute of Transportation Studies, University of California, Irvine.

Gourieroux, C., A. Monfort and E. Renault, 1987. "Generalized Residuals," *Journal of Econometrics*, 34, pp. 5-32.

Greene. W., 1981. "Sample Selection Bias As a Specification Error: Comment," *Econometrica*, 49, pp. 795-798.

Greene, W., 1990. *Econometric Analysis*, New York, Macmillan.

Greene, W., 1992. "A Statistical Model for Credit Scoring." Working Paper No. EC-92-29, New York University, Department of Economics, Stern School of Business.

Greene, W., 1994. "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," Working Paper 94-10, Department of Economics, Stern School of Business, New York University.

Greene, W., 1995. "Sample Selection in the Poisson Regression Model," Department of Economics, Stern School of Business, New York University, Working paper #95-06, 1995.

Greene, W. "Marginal Effects in the Bivariate Probit Model." Working Paper No. 96-11, Department of Economics, Stern School of Business, New York University, 1996

Greene, W., 2002. *LIMDEP Version 8.0, Reference Guide*, Plainview, NY, Econometric Software.

Greene, W., 2003. *Econometric Analysis, 5th Edition*, Englewood Cliffs, Prentice Hall.

Greene, W., 2004a. "Fixed Effects and Bias Due To The Incidental Parameters Problem in the Tobit Model," *Econometric Reviews*, 23, 2, pp. 125-147.

Greene, W., 2004b. "The Behavior of the Fixed Effects Estimator in Nonlinear Models," *The Econometrics Journal* , 7, 1, pp. 98-119.

Greene, W., 2004c. "Convenient Estimators for the Panel Probit Model." *Empirical Economics,* 29, 1, pp. 21–47.

Greene, W., 2005. "Functional form and Heterogeneity in Models for Count Data," *Foundations and Trends in Econometrics*, 1, 2, pp. 113-218.

Greene, W., 2006**. "**A General Approach to Incorporating Selectivity in a Model," Department of Economics, Stern School of Business, New York University, Working Paper 06-10.

Greene, W., 2007a. *LIMDEP Version 9.0: Reference Guide*, Plainview, New York, Econometric Software, Inc.

Greene, W., 2007b. *NLOGIT Version 4.0: Reference Guide*, Plainview, NY, Economertric Software.

Greene, W., 2008a. *Econometric Analysis, 6th Edition*, Englewood Cliffs, Prentice Hall.

Greene, W., 2008b. "A Stochastic Frontier Model with Correction for Selection," Department of Economics, Stern School of Business, New York University, Working Paper EC-08-09.

Greene, W., M. Harris, B. Hollingworth, P. Maitra, 2008. "A Bivariate Latent Class Correlated Generalized Ordered Probit Model with an Application to Modeling Observed Obesity Levels," Department of Economics, Stern School of Business, New York University, Working Paper 08-18.

Greene, W. and D. Hensher,, 2009. "Ordered Choices and Heterogeneity in Attribute Processing," *Journal of Transport Economics and Policy*, forthcoming..

Greene, W., L. Knapp and T. Seaks, 1993. "Estimating the Functional Form of the Independent Variables in Probit Models," *Applied Economics*, pp. 193-196.

Greenland, S., 1994. "Alternative Models for Ordinal Logistic Regression," *Statistics in Medicine*, 13, 1665-1677.

Greenwood, C. and V. Farewell, 1988. "A Comparison of Regression Models for Ordinal Data in Analysis of Transplanted Kidney Function," *Canadian Journal of Statistics*, 16, pp. 325-336.

Grizzle, J., C. Starmet and G. Koch, 1969. "Analysis of Categorical Data By Linear Models," *Biometrics*, 25, pp. 489-504.

Groot, W. and H. van den Brink, 1999. "Job Satisfaction with Preference Drift," *Economics Letters*, 63, 3, pp. 363-367.

Groot, W. and H. van den Brink, 2002. "Sympathy and the Value of Health," *Social Indicators Research*, 61, 1, pp. 97-120.

Groot, W. and H. van den Brink, 2003a. "Match Specific Gains to Marriage: A Random Effects Ordered Response Model," *Quality and Quantity*, 37, pp. 317-325.

Groot, W. and H. van den Brink, 2003b. "Firm-Related Training Tracks: A Random Effects Ordered Probit Model," *Economics of Education Review*, 22, 6, pp. 581-589.

Guo, J.Y., C.R. Bhat, and R.B. Copperman (2007) Effect of the Built Environment on Motorized and Non-Motorized Trip Making: Substitutive, Complementary, or Synergistic? *Transportation Research Record,* 2010, 1-11.

Gurland, J., J. Lee and P. Dahm, 1960. "Polychotomous Quantal Response in Biological Assay," *Biometrics*, 16, pp. 382-398.

Gupta, K, N. Kristensen and D. Pozzoli, 2008. "External Validation of the Use of Vignettes in Cross-Country Health Studies," Health Econometrics Workshop, Milan, December, Department of Economics, Aarhus School of Business, University of Aarhus.

Gupta, N., N. Kristersen and D. Pozzoli, 2008. "External Validation of Use of Vignettes in Cross-Country Health Studies," Department of Economics, Aarhus School of Business, University of Aarhus, Presented at the 2008 Health Econometrics Workshop, Milan.

Gustaffson, S. and Stafford, F., 1992. "Child Care Subsidies and Labor Supply in Sweden," *Journal of Human Resources,* 27, pp. 204-230.

Hafner, K., 2006. "And If You Liked the Movie, A Netflix Context May Reward You Handsomely," *New York Times*, October 12, 2006.

Halton, J.H. (1970) A Retrospective and Prospective Survey of the Monte Carlo Method. *SIAMReview*, 12, 1-63.

Hamermesch, D., 2004. "Subjective Outcomes in Economics," *Southern Economic Journal*, 71, 1, pp. 2-11.

Han, A. and J. Hausman, 1988. "Semiparametric Estimation of Duration and Competing Risk Models," Department of Economics, MIT, Manuscript.

Han, A. and J.A. Hausman (1990). Flexible parametric estimation of duration and competing risk models, Journal of Applied Econometrics, 5, 1-28.

Hahn, J. and G. Kuersteiner, 2003. "Bias Reduction for Dynamic Nonlinear Panel Data Models with Fixed Effects," Department of Economics, UCLA, Manuscript.

Hahn, J. and W. Newey, 2004. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 4, pp. 1295-1319.

Han, A. and J. Hausman, 1986. "Semiparametric Estimation of Duration and Competing Risk Models," Department of Economics, MIT, Working Paper 450.

Harvey, A., 1976. "Estimating Regression Models with Multiplicative Heteroscedasticity," *Econometrica*, 44, pp. 461-465.

Hardle, W., S. Huet, E. Mammen and E. Sperlich, 2004. "Bootstrap Inference in Semiparametric Generalized Additive Models," *Econometric Theory*, 20, pp. 265-300.

Hausman, J., 1978. "Specification Tests in Econometrics," *Econometrica*, 46, pp. 1251-1271.

Hausman, J., A. Lo and C. MacKinlay, 1992. "An Ordered Probit Analysis of Transaction Stock Prices," *Journal of Financial Economics*, 31, pp. 319-379.

Harris, M. and X. Zhao, 2007. "Modeling Tobacco Consumption with a Zero Inflated Ordered Probit Model," *Journal of Econometrics*, 141, pp. 1073-1099.

Hayashi, H., 2000. *Econometrics*, Princeton, Princeton University Press.

Heckman, J.J., 1978. "Dummy Endogenous Variables in a Simultaneous Equation System" *Econometrica* 46, 931–959.

Heckman, J., 1979. "Sample Selection Bias as a Specification Error," *Econometrica*, 47, pp. 153-161.

Heckman, J., 1981a. "Statistical Models for Discrete Panel Data," *In Structural Analysis of Discrete Data with Econometric Applications*, Edited by C. Manski and D. McFadden, MIT Press, Cambridge..

Heckman, J., 1981b. "Heterogeneity and State Dependence," in *Studies of Labor Markets*, edited by S. Rosen, NBER , University of Chicago Press, Chicago.

Heckman, J. J. and T. E. MaCurdy, 1981. "New Methods for Estimating Labor Supply Functions," in R. Ehrenberg, ed., *Research in Labor Economics*, Greenwich, CT., JAI Press, pp. 65–102.

Heckman, J. and S. Navarro, 2005. "A General Ordered Choice and Duration Model for Counterfactuals Motivated by Economic Analysis," Department of Economics, University of Chicago, Manuscript.

Heckman, J. and S. Navarro, 2007. "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 136, pp. 341-396.

Heckman, J. and B. Singer, 1984a. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models," *Econometrica*, 52, pp. 271-320.

Heckman, J. and B. Singer , 1984b. "Econometric Duration Analysis," *Journal of Econometrics*, 24, pp. 63-132.

Heckman, J. and MaCurdy, 1985. "A Simultaneous Equations Linear Probability Model," *Canadian Journal of Economics*, 18, pp. 28-37.

Heckman, J. and J. Snyder, 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators," *Rand Journal of Economics*, 28.

Heilbron, D., 1994. "Zero-Altered and Other Regression Models for Count Data with Added Zeros," *Biometrical Journal*, 36, pp. 531-547.

Hemmingsen, A., 1933. "The Accuracy of Insulin Assay on White Mice" *Quarterly Journal of Pharmacy and Pharmacology*, 6, 39-80 and 187-283.

Hensher, D., 2006. "How Do Respondents Process Stated Choice Experiments? – Attribute Consideration Under Varying Information Load," *Journal of Applied Econometrics*, 21, pp. 861-878.

Hensher, D. and Jones, S., 2007. Predicting Corporate Failure: Optimizing the Performance of the Mixed Logit Modeo, *ABACUS*, 43, 3, pp. 241-264.

Hensher, D., N. Smith, N. Milthorpe and P. Barnard, 1992. Dimensions of Automobile Demand: A longitudinal Study of Household Automobile Ownership and Use, Studies in Regional Science and Urban Economics, Elsevier Science Publishers, Amsterdam.

Hensher, D., J. Rose and W. Greene, 2005. *Applied Choice Analysis: A Primer*, Cambridge University Press, Cambridge.

Herbert, A., N. Gerry and N. McQueen, 2006. "A Common Genetic Variant is Associated with Adult and Childhood Obesity," *Science*, 312, pp. 279-283.

Hinde, J., G. Clarice and Demetrio, 1998., "Overdispersion in Models and Estimation," *Computational Statistics and Data Analysis*, 27, 2, pp. 151-170.

Holmes, M. and R. Williams, 1954. "The Distribution of Carriers of Streptococcus Pyogenes Among 2413 Healthy Children," *Journal of Hygiene*, 52, pp. 165-179.

Honore, B. and A. Lewbel, 2002. "Semiparametriic Binary Choice Panel Data Models without Strictly Exogenous Regressors," *Econometrica*, 70, pp. 2053-2063.

Hopkins, D. and G. King, 2008. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Comparability," Kennedy School of Government, Harvard University, Unpublished manuscript, http://gking.harvard.edu/files/implement.pdf

Horowitz, J., 1992. "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, pp. 505-531.

Horowitz, J., 1993. "Semiparametric Estimation of a Work-Trip Mode Choice Model." *Journal of Econometrics,* 58, pp. 49–70.

Hosmer, D and S. Lemeshow, 2000. *Applied Logistic Regression*, 2nd. Ed., New York, John Wiley and Sons.

Hsiao, C., 1986. *Analysis of Panel Data*, Cambridge, Cambridge University Press.

Hsiao, C., 2003. *Analysis of Panel Data, 2nd. Ed.*, Cambridge, Cambridge University Press.

Hutchison. V., 1985. "Ordinal Variable Regression Using the McCullagh (Proportional Odds) Model," *Canadian Journal of Statistics*, 16, pp. 325-336.

Hyslop, D., 1999. "State Dependence, Serial Correlation, and Heterogeneity in Labor Force Participation of Married Women," *Econometrica*, 67, 6, pp. 1255-1294.

Imai, K., G. King and O. Lau, 2008. "Zelig: Everyone's Statistical Sodtware," Department of Government, Harvard University. (http://gkng.harvard.edu/zelig/)

Inkmann, J., 2000. "Misspecified Heteroscedasticity in the Panel Probit Model: A Small Sample Comparison of GMM and SML Estimators," *Journal of Econometrics*, 97, 2, pp. 227-259.

ISI, 1982. "Citation Classic: Finney D. J. Probit Analysis," ISI, 31, August.

Jansen, J., 1990. "On the Statistical Analysis of Ordinal Data when Extravariation is Present," *Applied Statistics*, 39, pp. 75-84.

Jiminez , E. and B. Kugler, 1987. "The Earnings Impact of Training Duration in a Developing County: An Ordered Probit Selection Model of Colombia's Servicio Nacional de Aprendisaje," *Journal of Human Resources*, 22, 2, pp. 228.247.

Johnson, P., 1996. "A Test of the Normality Assumption in the Ordered Probit Model," *Metron*, 54, pp. 213-221.

Johnson, N., S. Kotz and A. Balakrishnan, 1994. *Continuous Univariate Distributions, Vol. I*, New York, John Wiley and Sons.

Johnson, V. and J. Abbot, 1999. *Ordinal Data Modeling*, New York, Springer-Verlag.

Jones, S. and D. Hensher, 2004. "Predicting Firm Financial Distress: A Mixed Logit Model," *The Accounting Review (American Accounting Association)*, 79, pp. 1011-1038.

Kadam, A and P. Lenk, 2008. "Bayesian Inference for Issuer Heterogeneity in Credit Ratings Migration" . *Journal of Banking and Finance*, Forthcoming. (SSRN:http://ssrn.com/abstract=1084006)

Kalbfleisch, J. and R. Prentice, 2002. *The Statistical Analysis of Failure Time Data, 2nd ed.*. New York, John Wiley and Sons.

Kalman, R., 1960. "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering (ASME Transactions)*, 82D, pp. 35-45.

Kao, C. and C. Wu, 1990. "Two Step Estimationi of Linear Models with Ordinal Unobserved Variables: The Case of Corporate Bonds," *Journal of Business and Economic Statistics*, 8, pp. 317-325.

Kapteyn, A., J. Smith and A. van Soest, 2007. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands," *American Economic Review*, 97, 1, pp. 461-473.

Kasteridis, P., M Munkin, and S. Yen., 2008. "A Binary-Ordered Probit Model of Cigarette Demand." *Applied Economics*, 41, forthcoming.

Katz, E., 2001. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation. *Political Analysis*, (4), pp. 379-384.

Kawakatsu, J. and A, Largey, 2009. "An EM Algorithm for Ordered Probit Models with Endogenous Regressors," *Econometrics Journal*, forthcoming.

Kay, R., and S. Little., 1986. "Assessing the Fit of the Logistic Model: A Case Study of Children with Haemolytic Uraemic Syndrome." *Applied Statistics,* 35, pp. 16–30.

Keele, L. and D. Park, 2005. "Difficult Choices: An Evaluation of Heterogeneous Choice Models," Presented at the 2004 Meeting of the American Political Science Association, Department of Politics and International Relations, Oxford University, Manuscript.

Kenny, L., L. Lee, G. Maddala and R. rost, 1979. "Returns to College Education: An Investigation of Self-Selection Bias and the Project Talent Data," *International Economic Review*, 20, 3, pp. 775-789.

Kerkhofs, M., Lindeboom, M., 1995. "Subjective Health measures and State Dependent Reporting Errors" *Health Economics* 4, pp. 221–235.

Kiefer, N., 1982. "Testing for Independence in Multivariate Probit Models." *Biometrika,* 69, pp. 161–166.

Kim, K., 1995. "A Bivariate Cumulative Probit Regression Model for Ordered Categorical Data," *Statistics in Medicine*, 14, pp. 1341-1352.

King, G., Murray, C. J., Salomon, J. A., and Tandon, A., 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research,"*American Political Science Review*, 98, 191–207, http://gking.harvard.edu/files/abs/vign-abs.shtml.

King, G. and Wand, J., 2007, "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes," *Political Analysis*, 15, 46–66, http://gking.harvard.edu/files/abs/cabs.shtml.

King, G. and J. Wand, 2007. "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes," *Political Analysis*, 15, pp. 46–66.

Kitamura,R. and D. Bunch (1989) Heterogeneity and State Dependence in Household Car Ownership: A Panel Analysis Using Ordered-Response Probit Models with Error Components. Research Report, UCD-TRG-RR-89-6, Transportation Research Group, University of California at Davis.

Kitamura, R., 1987. A Panel Analysis of Household Car Ownership and Mobility, Infrastructure Planning and Mangement, Proceedings of the Japan Society of Civil Engineers, 383/IV-7, pp. 13-27.

Kitamura, R., 1988. A Dynamic Model System of Household Car Ownership, Trip Generation and Modal Split, Model Development and Simulation Experiments, In Proceedings of the 14th Australian Road Research Board Conference, Part 3, Australian Road Research Board, Vermint South, Victoria, Australia, pp. 96-111.

Klein, R. and R. Sherman, 2002. "Shift Restrictions and Semiparametric estimation in Ordered Response Models," *Econometrica*, 70, pp. 663-692.

Klein, R. and R. Spady, 1993. "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, 61, pp. 387-421.

Knapp, L., and T. Seaks., 1992. "An Analysis of the Probability of Default on Federally Guaranteed Student Loans." *Review of Economics and Statistics,* 74, pp. 404–411.

Kohler, H. and J. Rodgers, 1999. "DF-Like Analyses of Binary, Ordered and Censored Variables using Probit and Tobit Approaches," *Behavior Genetics*, 20, 4, pp. 221-232.

Koop, G. and J. Tobias, 2006. "Semiparametric Bayesian Inference in Smooth Coefficient Models," *Journal of Econometrics*, 134, 1, pp. 283-315.

Krailo, M. and M. Pike, 1984. "Conditional Multivariate Logistic Analusis of Stratified Case-Control Studies," *Applied Statistics*, 44, 1, pp. 95-103.

Krinsky, I., and L. Robb, 1986. "On Approximating the Statistical Properties of Elasticities." *Review of Economics and Statistics,* 68, 4, pp. 715–719.

Krinsky, I., and L. Robb, 1990. "On Approximating the Statisticl Properties of Elasticities: Correction." *Review of Economics and Statistics,* 72, 1, pp. 189–190.

Krinsky, I., and L. Robb, 1991. "Three Methods for Calculating Statistical Properties for Elasticities." *Empirical Economics,* 16, pp. 1–11.

Kristensen, N. and E. Johansson, 2008. "New Evidence on Cross Country Differences in Job Satisfaction Using Anchoring Vignettes, *Labor Economics*, 15, 96-117.

Kuriama, K., Y. Kitibatake, Y. Oshima, 1998. "The Downward Bias Due to "No-Vote" Option in Contingent Valuation Study, World Congress of Environmental and Resource Economists,Venice. (http://www.f.waseda.jp/kkuri/research/workingpaper/WP9805.PDF) (1998)

Lancaster, T., 2000. "The Incidental Parameters Problem Since 1948," *Journal of Econometrics*, 95, pp. 391-413.

Lancaster, T., 2004. *An Introduction to Modern Bayesian Inference*, Oxford, Oxford University Press.

Lambert, D., 1992. "Zero-inflated Poisson Regression With An Application To Defects In Manufacturing," *Technometrics*, 34, 1, pp. 1-14.

Lawrence, C. and H. Palmer, 2002. Heuristics, Hillary Clinton and Health care Reform, Annual Meeting of the Midwest Political Science Association, Chicago.

Lechner, M., 1991. "Testing Logit Models in Practice," *Empirical Economics*, 16, pp. 77-108.

Lee, L. (1983) Generalized Econometric Models with Selectivity, Econometrica, 51, 507-512.

Lee, L. and R. Trost, 1978. Estimation of Some Limited Dependent Variable Models with Application to Housing Demand," *Journal of Econometrics*, 8, pp. 357-382.

Lewbel, A., 1997. "Semiparametric Estimation oof Location and Other Discrete Choice Moments," *Econometric Theory,* 13, pp. 32-51.

Lewbel, A. and S. Schennach, 2007. "A Simple Ordered Data Estimator for Inverse Density Weighted Expectations," *Journal of Econometrics*, 136, pp.189-211..

Lewbel, A., 2000. Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, pp. 145-177.

Li, M. and J. Tobias, 2006a. "Calculus Attainment and Grades Received in Intermediate Economic Theory," *Journal of Applied Econometrics*, 21,6, pp. 893-896.

Li, M. and J. Tobias 2006b. "Bayesian Analysis of Structural Effects in an Ordered Equation System," Department of Economics, Iowa State University, Working Paper.

Li, M. and J. Tobias 2006c. "Bayesian Analysis of Structural Effects in An Ordered Equation System," *Studies in Nonlinear Dynamics & Econometrics*, 10, 4, Article 7, pp. 1-24.

Li, Q and J. Racine, 2007. *Nonparametric Econometrics*, Princeton University Press, Princeton.

Lillard, L. and E. King, 1987. "Education Policy and Schooling Attainment in Malaysia and the Philippines," *Economics of Education Review*," 6, pp. 167-181.

Lindeboom, M. and E. van Doorslayer, 2003. "Cut Point Shift and Index Shift in Self Reported Health," *Ecuity III Project* Working Paper #2.

Long, S. 1997. *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA, Sage Publications.

Long, S. and J. Freese, 2006. *Regression Models for Categorical Dependent Variables Using Stata*, College Station, Stata Press.

Machin, S. and A. Vignoles, 2005. *What's the Good of Education? The Economics of Education in the UK*, Princeton, N.J., Princeton University Press.

MacKinnon, J., 1992. "Model Specification Tests and Artificial Regressions," *Journal of Economic Literature*, 30, 1, pp. 102-146.

Maddala, J., 1983. *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge, Cambridge University Press.

Maddala, G., 1987. "Limited Dependent Variable Models Using Panel Data," *Journal of Human Resources*, 22, pp. 307-338.

Magee, L., J. Burbidge and A. Robb, 2000. "The Correlation Between Husband's and Wife's Education: Canada , 1971-1996," SEDAP Research Paper #24, McMaster.

Mannering, F. and Winston, C. (1985) A Dynamic Analysis of Household Vehicle Ownership and Utilization, Rand Journal of Economics, 16, 215-236.

Manski, C., 1975. "The Maximum Score Estimator of the Stochastic Utility Model fo Choice," *Journal of Econometrics*, 3, pp. 205-228.

Manski, C., 1985. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, pp. 313-333.

Manski, C., 1985. *Identification Problems in the Social Sciences.* Cambridge: Harvard University Press.

Manski, C., 1987. "Semiparametric Analysis of the Random Effects Linear Model from Binary Response Data," *Econometrica*, 55, pp. 357-362.

Manski, C., 1986. "Operational Characteristics of the Maximum Score Estimator," *Journal of Econometrics*, 32, pp. 85-100.

Manski, C., 1988. "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, pp. 729-738.

Manski, C., and S. Lerman, 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica,* 45, pp. 1977–1988.

Manski, C. and S. and Thompson, 1985. "MSCORE: A Program for Maximum Score Estimation of Linear Quantile Regressions from Binary Response Data, Mimeo, Department of Economics, University of Wisconsin, Madison.

Marcus, A. and W. Greene, 1983. "The Determinants of Rating Assignment and Performance," Working Paper CRC528, Alexandria, VA, Center for Naval Analyses.

Matzkin, R., 1993. "Nonparametric Identification and Estimation of Polytomous Choice Models." *Journal of Econometrics,* 58, pp. 137–168.

McCullagh, P., 1977. "Analysis of Ordered Categorical Data," Ph.D. Thesis, University of London.

McCullagh, P., 1979, "The Use of the Logistic Function in the Analysis of Ordinal Data," *Bulletin of the International Statistical Institute*, 48, pp. 21-33.

McCullagh, P., 1980. "Regression Models for Ordered Data," *Journal of the Royal Statistical Society, Series B (Methodological)*, 42, pp. 109-142.

McCullagh, P. and J. Nelder, 1983. *Generalized Linear Models,* London, Chapman and Hall.

McCullagh, P. and J. Nelder, 1989. *Generalized Linear Models, 2nd. Ed.*, London, Chapman and Hall.

McElvey, R. and W. Zavoina, 1971. "An IBM Fortran IV Program to Perform N-Chotomus Multivariate Probit Analuysis," *Behavioral Science*, 16, 2, March, pp. 186-187.

McElvey, R. and W. Zavoina, 1975. "A Statistical Model for the Analysis of Ordered Level Dependent Variables," *Journal of Mathematical Sociology*, 4, pp. 103-120.

McFadden, D., 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics*, New York, Academic Press.

McFadden, D. and K. Train, 2000. "Mixed MNL Models for Discrete Choice," *Journal of Applied Econometrics*, 15, pp. 447-470.

McKenzie, C., 1998. "Microfit 4.0," *Journal of Applied Econometrics*, 13, pp. 77-90.

McLachlan, G. and D. Peel, 2000. *Finite Mixture Models*, New York, John Wiley and Sons.

McLachlan, G. and D. Peel, 2000. *Finite Mixture Models*, New York, John Wiley and Sons.

McVicar, M. and J. McKee, 2002. "Part Time Work During Post-Compulsory Education and Examination Performance: Help or Hindrance" *Scottish Journal of Political Economy*, 49,4, pp. 393-406.

Mehta, C. and N. Patel, 1995. "Exact Logistic Regression: Theory and Examples," *Statistics in Medicine*, 14, pp. 2143-2160.

Melenberg, B., and A. van Soest, 1996. "Parametric and Semi-Parametric Modelling of Vacation Expenditures." *Journal of Applied Econometrics,* 11, 1, pp. 59–76.

Mitchell, J. and M. Weale, 2007. "The Rationality and Reliability of ExpectationsReported by British Households: Micro Evidence From the BHPS, National Institute of Economic and Social Research, Manuscript. (http://www.niesr.ac.uk/pubs/dps/DP287.pdf)

Mohanty, M., 2002. "A Bivariate Probit Approach to the Determination of Employment: A Study of Teen Employment Differentials in Los Angeles County," *Applied Economics,* 34, 2, pp. 143–156.

Mora, J. and A. Moro-Egido, 2008. "On Specification Testing of Ordered Probit Models," *Journal of Econometrics*, 143, pp. 292-205.

Mora, N., 2006., "Sovereign Credit Ratings: Guilty Beyond Reasonable Doubt?" *Journal of Banking and Finance*, 30, pp. 2041-2062.

Mora, J. and A. Moro-Egido, 2008. "On Specification Testing of Ordered Discrete Choice Models," *Journal of Econometrics*, 143, pp. 191-205.

Mullahy, J., 1986. "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, pp. 341-365.

Mullahy, J., 1997. "Heterogeneity, Excess Zeros and the Structure of Count Data Models," *Journal of Applied Econometrics*, 12, pp. 337-350.

Müller, G. and C. Czado, 2005. "An Autoregressive Ordered Probit Model with Application to High Frequency Finance," *Journal of Computational and Graphical Statistics*, 14, pp. 320-338.

Mundlak, Y., 1978. "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 56, pp. 69-86.

Munkin, M. and P. Trivedi, 2008. "Bayesian Analysis of the Ordered Probit Model with Endogenous Selection," *Journal of Econometrics*, 143, pp. 334-348.

Murad, H., A. Fleischman, S. Sadetzki, O. Geyer and L. Freedman, 2003. "Small Samples and Ordered Logistic Regression: Does it Ielp to Collapse Categories of Outcome?" *The American Statistician*, 57, 3, pp. 155-160.

Murphy, A., 1994. "Artificial Regression Based LM tests of Misspecification for Discrete Choice Models," *Economic and Social Review*, 26, pp. 69-74.

Murphy, A, 1996. "Simple LM Tests of Misspecification for Ordered Logit Models," *Economics Letters*, 52, pp. 137-141.

Murphy, K. and R. Topel, 2002. "Estimation and Inference in Two Stem Econometric Models," *Journal of Business and Economic Statistics*, 20, pp. 88-97 (reprinted from 2, pp. 370-379).

Murray, C., A. Tandon, C. Mathers, R. Sudana, 2002. "New Approaches to Enhance Cross-Population Comparability of Survey Results, in Murray, C., A. Tandon, R. Mathers, R. Sudana, eds., *Summary Measures of Population Health*, Ch. 8.3 World Health Organization.

Nagin, D. and J. Waldfogel, 1995. "The Effect of Criminality and Conviction on the Labor Market Status of Young British Offenders," *International Review of Law and Economics*, 15, 1, pp. 109-126.

Nagler, J., 1994. "Scobit: An Alternative Estimator to Logit and Probit," *American Journal of Political Science*, 38, 1, pp. 230-255.

Nakosteen, R. and M. Zimmer, 1980. "Migration and Income: The Question of Self Selection," *Southern Economic Journal*, 46, pp. 840-851.

NDSHS, 2001. Computer Files for the Unit Record Data from the National Drug Strategy Household Surveys.

Nelder, J. and R. Wedderburn, 1972. "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, pp. 370-384.

Newey, W., 1985. "Maximum Likelihood Specification and Testing and Conditional Moment tests," *Econometrica*, 53, pp. 1047-1070.

Newey, W., 1987. "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables." *Journal of Econometrics,* 36, pp. 231–250.

Nerlove, M. and J. Press, 1972. "Univeriate and Multivariate Log-Linear and Logistic Models," Santa Monica, CA, RAND – R1306-EDA/HIS.

Neyman, J. and E. Scott, 1948. "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, pp. 1-32.

Norton, E. and C. Ai, 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters,* 80, 1, pp. 123−129.

Olsen, R. 1978. "A Note on the Uniqueness of the Maximum Likelihood Estimator in the Tobit Model," *Econometrica*, 46, pp. 37-44.

Olssen, U., 1979. "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 4, pp. 443-460.

Olsson, U., 1980. "Measuring of Correlation in Ordered Two Way Contingency Tables," *Journal of Marketing Research*, 17, 3, pp. 391-394

Orme, C., 1990. "The small-sample performance of the information-matrix test," Journal of Econometrics 46, pp. 309–331

Pagan, A. and F. Vella, 1989. "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 4, Supplement, pp. S29-S59.

Paternoster, R. and R. Brame, 1998. "The Structural Similarity of Processes Generating Criminal and Analogous Behaviors." *Criminology* 36, pp. :633-670.

Pearson, K., 1914. *Tables for Statisticians and Biometricians: Part I*, Cambridge, Cambridge University Press.

Plackett, R., 1974. *The Analysis of Categorical Data*, London, Griffin.

Plackett, R., 1981. *The Analysis of Categorical Data, 2nd. Ed.*, London, Griffin

Popuri, Y.D., and C.R. Bhat (2003) On Modeling Choice and Frequency of Home-Based Telecommuting. *Transportation Research Record*, 1858, 55-60.

Pratt, J., 1981. "Concavity of the Log Likelihood," *Journal of the American Statistical Association*, 76, pp. 103-116.

Pregibon, D., 1984. "Book Review: P. McCullagh and J. A. Nelder, *Generalized Linear Models*," *The Annals of Statistics*, 12, 4, pp. 1589-1596.

Prescott, E. and M. Visscher, 1977. "Sequential Location among Firms with Foresight," *Bell Journal of Economics*, 8, pp. 378–893.

Pudney,S. and M. Shields, 2000. "Gender, Race, Pay and Promotion in the British Nursing Profession: Estimation of a Generalized Ordered Probit Model," *Journal of Applied Econometrics*, 15, pp. 367-399.

Purvis, L. (1994) Using Census Public Use Micro Data Sample to Estimate Demographic and Automobile Ownership Models, Transportaton Research Record, 1443, 21-30.

Quednau, H., 1988. "An Extended Threshold Model for Analyzing Ordered Categorical Data," *Biometrical Journal*, 31, pp. 781-793.

Rabe-Hesketh, S**.**, A. Skrondal, A. and A. Pickles, 2005. "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics,* 128, pp. 301-323.

Rasch, G., 1960. "Probabilistic Models for Some Intelligence and Attainment Tests," Copenhagen, Danish Institute for Educational Research.

Raudenbusch, S. and A. Bryk, 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Thousand Oaks, CA.

Ridder, G., 1990. "The Non-parametric Identification of Generalized Accelerated Failure-Time Models," *Review of Economic Studies*, 57 pp. 167–181.

Riphahn, R., A. Wambach and A. Million, 2003. "Incentive Effects on the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, 18, 4, pp. 387-405.

Ronning, G., 1990. "The Informational Content of Responses from Business Surveys," In J. Florens, M. Ivaldi, J. Laffont and F. Laisney, eds., *Microeconometrics: Surveys and Applications*, Oxford, Blackwell.

Ronning, G. and M. Kukuk, 1996. "Efficient Estimation of Ordered Probit Models," *Journal of the American Statistical Association*, 91, 435, pp. 1120-1129.

Rosetti, C., 2008, "Ordered Probit Models with Anchoring Vignettes," Manuscript, Faculty of Economics, University of Rome.

Ruud, P., 1986. "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of the Distribution." *Journal of Econometrics,* 32, pp. 157–187.

Sajaia, Z., 2008. "Maximum Likelihood Estimation of a Bivariate Ordered Probit Model: Implementation and Monte Carlo Simulations," *The Stata Journal*, 4, 2, pp. 1-18.

Sanko, S., H. Maesoba, D. Dissanayake, T. Yamamoto, and T. Morikawa, 2004. "Inter-temporal and Inter-Regional Analysis of Household Cars and Motorcycles Ownership Behaviours in Asian Big Cities," Graduate School of Environmental Studies, Nagoya University, manuscript. (http://cost355.inrets.fr/IMG/doc/SAKURA_Yamamoto_.doc)

SAS Institute, 2008. *SAS/STAT User's Guide. Version 9.2*, Cary NC, SAS Institute.

SAS, 2008. *SAS User's Guide*, Cary NC, SAS.

Scott, D. and K. Axhausen, 2006. "Household Mobility Tool Ownership, Modeling Interactions Between Cars and Season Tickets," *Transportation*, 33, 4, pp. 311-328.

Scott, D. and Kanaroglou, P., 2001. "An Activity-Episode Generation Model that Captures Interactions Between Household Heads: Development and Empirical Analysis," *Transportation Research B: Methodological*, 36, 10, pp. 875-896

Scotti, C., 2006. "A Bivariate Model of Fed and ECB Main Policy Rates," Board of Governors, Internationall Finance Discussion paper 875

Severini, T. and J. Staniswalis, 1994. "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, pp. 501-511.

Shaked, A. and J. Sutton, 1982. "Relaxing Price Competition through Product Differentiation," *Review of Economic Studies*, 49, pp. 3–13.

Simonoff, J., 2003. *Analyzing Categorical Data*, New York, Springer.

Sirven, N., B. Santos-Egglmann, J. Spagnoli, 2008, "Comparability of Health Care Responsiveness in Europe Using Anchoring Vignettes," Working Paper 15, IRDES (France).

Smith, R., 1989. "On the Use of Distributional Misspecification Checks in Limited Dependent Variables," *Economic Journal*, 99, pp. 178-192.

Snell, E., 1964. "A Scaling Procedure for Ordered Categorical Data," *Biometrics*, 20, pp. 592-607.

Stata, 2008. *Stata, Version 8.0*, College Station TX, Stata Corp.

Stewart, M., 1983. "On Least Squares Estimation When the Dependent Variable Is Grouped," *Review of Economic Studies*, 50, pp. 141-149.

Stewart, M. 2003. "Semi-nonparametric Estimation of Extended Ordered Probit Models," Department of Economics, University of Warwick, Manuscript.

Stewart, M., 2005. "A Comparison of Semiparametric Estimators for the Ordered Response Model," *Computational Statistics and Data Analysis*, 49, pp. 555-573.

Stokes, H., 2004**.** "On the Advantage of UsingTwo or More Econometric Software Systems to Solve the Same Problem." *Journal of Economic and Social Measurement*, 29, 307-320.

Stutzer, A. and R. Lalive, 2001. "The Role of Social Work Norms in Job Searching and Subjective Well-Being," IZA Discussion Paper Number 300, Institute for the Study of Labor.

Swamy, P., 1970. "Efficient Inference in a Random Coefficient Regerssion Model," *Econometrica*, 38, pp. 311-323.

Swamy, P., 1971. *Statistical Inference in Random Coefficient Regression Models*, New York, Springer-Verlag.

Tandon, A., C. Murray, J. Aalamon, G. King, 2004, "Statistical Models for Enhancing Cross-Population Comparability," Ch. 55 in Murray and Evans, Eds., *Health System Performance Assessment: Debates, Methods and Empiricism*, World Health Organization.

Tanner, J. and W. Wong, 1987. "The Calculation of Posterior Distributions byData Augmentation," *Journal of the American Statistical Association*, 82, pp. 528-549.

Tattersfield, F., C. Gimingham and H. Morris, 1925. "Studies on Contact Insecticides. Part III. A Quantitative Examination of the Insecticidal Action of Chlor-, Nitro- and Hydroxyl Derivatives of Benzene and Naphthalene," *Annals of Applied Biology*, 12, p. 218.

Tauchen, G., 1985. "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, pp. 415-443.

Terza, J., 1983. "A Two Stage Estimator for Models with Ordinal Selectivity," *Proceedings of the Business and Economics Section of the American Statistical Association*, pp. 484-486.

Terza, J., 1985. "Ordered Probit: A Generalization," *Communications in Statistics – A. Theory and Methods*, 14, pp. 1-11.

Terza, J., 1987. "Estimating Linear Models with Ordinal Qualitative Regressors," *Journal of Econometrics*, 34, pp. 275-291.

Terza, J., 1998. "Estimating Count Data Models with Endogenous Switching and Endogenous Treatment Effects," *Journal of Econometrics*, 84, pp. 129-154.

Theil, H., 1969. "A Multinomial Extension of the Linear Logit Model," International Economic Review, 10, pp. 251-259.

Theil, H., 1970. "On the Estimation of Relationships Involving Qualitative Variables," *American Journal of Sociology*, 76, pp. 103-154.

Theil, H., 1971. *Principles of Econometrics*, New York, John Wiley and Sons.

Thompson, C., 2008. "If You Liked This, You're Sure To Love That," *New York Times Magazine*, November 23, 2008, pp. 74-79.

Tobin, J., 1958. "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, pp. 24-36.

Tomoyuki, F. and F. Akira, 2006. "A Quantitative Anaysis on Tourists' Consumer Satisfaction Via the Bayesian Ordered Probit Model," Journal of the City Planning Institute of Japan, 41, pp. 2-10. (In Japanese)

Train, K. (1986) Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand, MIT Press

Train, K., 2003. *Discrete Choice Methods with Simulation*, Cambridge, Cambridge University Press.

Trost, R. and L. Lee, 1978. "Technical Training and Earnings: A Polychotomous Choice Model with Selectivity," *Review of Economics and Statistics*, 66, 1, pp. 151-156.

Tsay, R., 2002. *Analysis of Financial Time Series,* New York, John Wiley and Sons.

Tsay, R., 2005. *Analysis of Financial Time Series, 2nd Ed.*, New York, John Wiley and Sons.

Tutz, G., 1989. "Compound Regression Models for Categorical Ordinal Data," *Biometrical Journal*, 31, pp. 259-272.

Tutz, G., 1990. "Sequential Item Response Models with an Ordered Response," *British Journal of Mathematical and Statistical Psychology*, 43, pp. 39-55.

Tutz, G., 1991. "Sequential Models in Ordered Regression," *Computational Statistics and Data Analysis*, 11, pp. 275-295.

TSP, 2005. *TSP Version 5.0*, TSP International, Palo Alto, CA.

UCLA/ATS, 2008. Academic Technology services

(http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm)

Uebersax, J., 1999. "Probit Latent Class Analysis with Dichotomous or Ordered Category Measures: Conditional Independence/Dependence Models," *Applied Psychological Measurement*, 23, pp. 283-297.

Umyarov, A. and A. Tuzhlin, 2008. "Iproving Collaborative Filtering Recommendations Using External Data," *Proceedings of the IEEE ICDM 2008 Conference*.

van Soest, A., L. Delaney, C. Harmon, A. Kapteyn and J. Smith, 2007. "Validating the Use of Vignettes for Subjective Threshold Scales," Working Paper WP/14/2007, Geary Institute, University College, Dublin.

Veall, M. and K. Zimmermann, 1992. "Pseudo-R^2 in the Ordinal Probit Model." *Journal of Mathematical Sociology,* 16, pp. 333–342.

Vella, F., 1998. "Estimating Models with Sample Selection Bias," *Journal of Human Resources*, 33, pp. 127-170.

Verbeek, M., 1990. "On the Estimation of a Fixed Effects Model with Selectivity Bias," *Economics Letters*, 34, pp. 267-270.

Verbeek, M. and T. Nijman, 1992. "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 33, pp. 681-703.

Vuong, Q., 1989. "Likelihood Ratio Tests for Model Selection and Nonnested Hypotheses," *Econometrica*, 57, 2, pp. 307-333.

Vytlacil, E., 2006. "Ordered Discrete Choice Selection Models and Local Average Treatment Effect Assumptions: Equivalence, Nonequivalence and Representation Results," *Review of Economics and Statistics*, 88, pp. 578-581.

Walker, S. and D. Duncan, 1967. "Estimation of the Probability of an Event As a Function of Several Independent Variables," *Biometrika*, 54, pp. 167-179.

Wand, J., G. King and O. Lau, O. 2008. "Anchors: Software for Anchoring Vignettes Data," *Journal of Statistical Software*.

Wedderburn, A., 1974. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, pp. 439-447.

Weiss, A., 1993. "A Bivariate Ordered Probit Model with Truncation: Helmet Use and Motorcycle Injuries," *Applied Statistics*, 42, pp. 487-99.

Weiss, A., 1997. "Specification tests in Ordered Logit and Probit Models," *Econometric Reviews*, 16, 4, pp. 361-391.

White, H., 1980. "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48, pp. 87-838.

White, H., 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica,* 53, pp. 1–16.

Williams, R., 2006. "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables," Department of Sociology University of Notre Dame, Notre Dame, IN, *The Stata Journal*, 6, 1, pp. 58–82.

Willis, J., 2006. "Magazine Prices Revisited," *Journal of Applied Econometrics*, 21, 3, pp. 337-344.

Windmeijer, F., 1995. "Goodness of Fit Measures in Binary Choice Models." *Econometric Reviews,* 14, pp. 101–116.

Winkelmann, R., 2005. "Subjective Well-being and the Family: Results from an Ordered Probit Model with Multiple Random Effects" *Empirical Economics*, 30, 3, pp 749-761,

Winkelmann, L. and R. Winkelmann, 1998. "Why are the Unemployed So Unhappy? Evidence from Panel Data," *Economica*, 65, pp. 1-15.

Winship, C. and R. Mare, 1984. "Regression Models with Ordered Variables," *American Sociological Review*, 49, 512-525.

Wooldridge, J., 2002a. "Inverse Probability Weighted M Estimators for Sample Stratification, Attrition and Stratification," *Portuguese Economic Journal*, 1, pp. 117-139.

Wooldridge, J., 2002b. *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MIT Press.

Wooldridge, J., 2008. "Cluster-Sample Methods in Applied Econometrics: Aa Extended Analysis," Department of Economics, Michigan State University, Unpublished.

Wu, D., 1973. "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," *Econometrica*, 41, pp. 733-750.

Wynand, P. and B. van Praag, 1981. "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics*, 17, pp. 229-252.

Zabel, J., 1992. "Estimating Fixed and Random Effects Models with Selectivity," *Economics Letters*, 40, pp. 269-272.

Zavoina, W. and R. McKelvey, 1969, "A Statistical Model for the Analysis of Legislative voting Behavior," Presented at the meeting of the American Political Science Association.

Zhang, J., 2007. "Ordered Probit Modeling of User Perceptions of Protected Left-Turn Signals," *Journal of Transportation Engineering*., 133, 3, pp. 205-214.

Zhang, Y, F. Liang and Y. Yuanchang, 2007. "Crash Injury Severity Analysis Using a Bayesian Ordered Probit Model," Transportation Research Board, Annual Meeting, Paper Number 07-2335.

Zigante, V., 2007. "Ever Rising Expectations – The Determinants of Subjective Welfare in Croatia," School of Economics and Management, Lund University, Masters Thesis (www.essays.se/about/Ordered+Probit+Model/).

Index