

# Discrete Choice Modeling

*William Greene\**

## Abstract

We detail the basic theory for models of discrete choice. This encompasses methods of estimation and analysis of models with discrete dependent variables. Entry level theory is presented for the practitioner. We then describe a few of the recent, frontier developments in theory and practice.

## Contents

- 0.1 Introduction
- 0.2 Specification, estimation and inference for discrete choice models
  - 0.2.1 Discrete choice models and discrete dependent variables
  - 0.2.2 Estimation and inference
  - 0.2.3 Applications
- 0.3 Binary Choice
  - 0.3.1 Regression models
  - 0.3.2 Estimation and inference in parametric binary choice models
  - 0.3.3 A Bayesian estimator
  - 0.3.4 Semiparametric models
  - 0.3.5 Endogenous right hand side variables
  - 0.3.6 Panel data models
  - 0.3.7 Application
- 0.4 Bivariate and multivariate binary choice
  - 0.4.1 Bivariate binary choice
  - 0.4.2 Recursive simultaneous equations
  - 0.4.3 Sample selection in a bivariate probit model
  - 0.4.4 Multivariate binary choice and the panel probit model
  - 0.4.5 Application
- 0.5 Ordered choice
  - 0.5.1 Specification analysis
  - 0.5.2 Bivariate ordered probit models
  - 0.5.3 Panel data applications
  - 0.5.4 Applications
- 0.6 Models for counts
  - 0.6.1 Heterogeneity and the negative binomial model
  - 0.6.2 Extended models for counts: Two part, zero inflation, sample selection, bivariate
  - 0.6.3 Panel data models
  - 0.6.4 Applications
- 0.7 Multinomial unordered choices
  - 0.7.1 Multinomial logit and multinomial probit models
  - 0.7.2 Nested logit models
  - 0.7.3 Mixed logit and error components models
  - 0.7.4 Applications
- 0.8 Summary and Conclusions

To appear as “Discrete Choice Modeling,” in *The Handbook of Econometrics: Vol. 2, Applied Econometrics*, Part 4.2., ed. T. Mills and K. Patterson, Palgrave, London, forthcoming, 2008.

\* Department of Economics, Stern School of Business, New York University, 44 West 4<sup>th</sup> St., 7-78, New York, NY 10012, <http://www.stern.nyu.edu/~wgreene>, [wgreene@stern.nyu.edu](mailto:wgreene@stern.nyu.edu).

## 0.1 Introduction

This review will survey models for outcomes that arise through measurement of discrete consumer choices, such as whether to vote for a particular candidate, whether to purchase a car, how to get to work, whether to purchase insurance, where to shop or whether to rent or buy a home or a car. Traditional economic theory for consumer choice – focused on utility maximization over bundles of continuous commodities – is relatively quiet on the subject of discrete choice among a set of specific alternatives. Econometric theory and applications, in contrast, contain a vast array of analyses of discrete outcomes; *discrete choice modeling* has been one of the most fruitful areas of study in econometrics for several decades. There is a useful commonality in much of this. One can build an overview of models for discrete outcomes on a platform of individual maximizing behavior. Given that the literature is as vast as it is, and we have but a small number of pages within which to package it, this seems like a useful approach. In what follows, we will survey some of the techniques used to analyze individual *random utility* maximizing behavior.

We emphasize that we have chosen to focus on models for discrete *choice*, rather than models for discrete *dependent variables*. This provides us with several opportunities to focus and narrow this review. First, it allows us to limit the scope of the survey to a reasonably manageable few types of models. As noted, the literature on this topic is vast. We will use this framework to select a few classes of models that are employed by analysts of individual choice. It also gives us license to make a few major omissions that might otherwise fall under the umbrella of discrete outcomes. One conspicuous case will be models for counts. Event counts are obviously discrete – models for them are used to study, e.g., traffic incidents, incidence of disease, health care system utilization, credit and financial markets, and an array of other settings. Models for counts can occupy an entire library of its own in this area – two excellent references are Cameron and Trivedi (1998) and Winkelmann (2003). But, this area will extend far beyond our reach. On the other hand, applications in health economics (system utilization) and industrial organization (patents and innovations), do lead to some settings in which individual or firm choice produces a count response. We will briefly consider models for counts from this standpoint. The reader will no doubt note other areas of discrete response analysis that are certainly important. Space limitations force us to consider a fairly small number of cases.

The study proceeds as follows: Section 2 will detail the estimation and inference tools used throughout the remainder of the survey, including the basic results in maximum likelihood estimation. Section 3 will analyze in detail the fundamental pillar of analysis of discrete choice, the model for binary choice – that is the choice between two alternatives. Most of the applications that follow are obtained by extending or building on the basic binary choice model. Thus, we will examine the binary choice model in greater detail than the others, as it also provides a convenient setting in which to develop the estimation and inference concepts that carry over to the other models as well. Section 4 considers the immediate extension of the binary choice, bivariate and multivariate binary choice models. Section 5 returns to the single choice setting, and examines ordered choice models. Models for count data are examined in Section 6. Each of the model classes mentioned has been analyzed using cross sections and panel data. The application of familiar panel data methods to discrete choices model is described in each section. Rather than consider the panel data version of each model separately, we have gathered several common results and features in a single section. Finally, Section 7 turns to an area of literature in its own right, multinomial choice modeling. As before, but even more so here, we face the problem of surveying a huge literature in a few pages. Section 8 will describe the most fundamental elements of multinomial choice analysis, and point the reader toward some more detailed sources in the literature. Some conclusions are drawn in Section 8.

## 0.2 Specification, estimation and inference for discrete choice models

The classical theory of consumer behavior provides the departure point for economic models of discrete individual choice.<sup>1</sup> A representative consumer with preferences represented by a utility function defined over the consumption of a vector of goods,  $U(\mathbf{d})$ , is assumed to maximize this utility subject to a budget constraint,  $\mathbf{x}'\mathbf{d} \leq I$ , where  $\mathbf{x}$  is a vector of prices and  $I$  is income (or total expenditure). Assuming the necessary continuity and curvature conditions, a complete set of demand equations,  $\mathbf{d}^* = \mathbf{d}(\mathbf{x}, I)$  results.<sup>2</sup> To extend the model of individual choice to observed market data, the demand system is assumed to hold at the aggregate level, and random elements (disturbances) are introduced to account for measurement error or optimization errors.

Since the 1960s, the availability of survey data on individual behavior has obviated the heroic assumption underlying the aggregate utility function or the (perhaps slightly less heroic) assumptions underlying the aggregate demand system. That progression has evolved to the contemporary literature with the appearance of large, detailed, high quality panel surveys such as the German Socio-Economic Panel Survey [GSOEP, see Hujer and Schneider (1989)] that we will use in this study and the British Household Panel Survey (BHPS, [www.iser.essex.ac.uk/ulsc/bhps](http://www.iser.essex.ac.uk/ulsc/bhps)) to name only two of many. The analysis of individual data to which the original theory applies has called for (at least) two more detailed developments of that theory.

First, the classical theory has relatively little to say about the *discrete* choices that consumers make. Individual data detail career choices, voting preferences, travel mode choices, discretized measures of the strength of preferences, participation decisions of all sorts such as labor supply behavior, whether to make a large purchase, whether to migrate, and on and on. The classical, calculus based theory of decisions made at the margins of consumption will comment on, for example, how a large a refrigerator a consumer will buy, but not whether they will buy a refrigerator instead of a car (this year), or what brand of refrigerator or car they will choose.

Second, the introduction of random elements in models of choice behavior as *disturbances*, is much less comfortable at the individual level than in market demands. Researchers have considered more carefully the appropriate sources and form of random variation in individual models of discrete choice.

The *random utility model* of discrete choice provides the most general platform for the analysis of discrete choice. The extension of the classical theory of utility maximization to the choice among multiple discrete alternatives provides a straightforward framework for analyzing discrete choice in probabilistic, statistical, ultimately econometric terms.

### 0.2.1 Discrete choice models and discrete dependent variables

Denote by ' $i$ ' a consumer who is making a choice among a set of  $J_{it}$  choices in choice situation  $t$ . To put this in a context which will help to secure the notation, envision a *stated choice experiment* in which individual  $i$  is offered the choice of several,  $J_{i1}$ , brands of automobiles with differing prices and characteristics and asked which they most prefer. In a second round of the experiment, the interviewer changes some of the features of some of the cars, and repeats the question. Denote by  $A_{i1}, \dots, A_{i, J_{it}}$ ,  $J_{it} \geq 2$ , the set of alternatives available to the individual in choice situation  $t$ . It will be convenient to adopt the panel data notation, in which ' $t$ ' denotes '*time*.' The generality of the notation allows the choice set to vary from one individual to another, and across choice situations for the same individual. In most of what follows, we will not need this level of generality, but the models to be developed will accommodate it.

---

<sup>1</sup> For a lengthy and detailed development of these ideas, see Daniel McFadden's Nobel Prize Lecture, McFadden (2001).

<sup>2</sup> See, as well, Samuelson (1947) and Goldberger (1987).

We will formulate a model that describes the consumer choice in probabilistic terms. (A bit more of the underlying behavioral theory is presented in Section 0.7.) The ‘model’ will consist of a probability distribution defined over the set of choices,

$$P_{it,j} = \text{Prob}(\text{consumer } i \text{ makes choice } j \text{ at time } t \mid \text{choice set}), j = 1, \dots, J_{it}.$$

The manner in which the probabilities arise is an essential feature of the model. As noted earlier, choices are dependent on the environment in which they are made, which we characterized in terms of income,  $I$ , and prices,  $\mathbf{x}$ . Individual heterogeneity may be measured by such indicators as family size, gender, location, and so on, which we collect in a set of variables,  $\mathbf{z}$ , and unmeasured and therefore random from the point of view of the analyst, which we denote as  $\mathbf{u}$ . Common elements of the choice mechanism that constitute the interesting quantities that the analyst seeks to draw statistical inference about will be parameters,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\gamma}$ , and so on.<sup>3</sup> For purposes of translating the underlying choice process into an estimable econometric model, we define the choice indicators,

$$d_{it,j} = 1 \text{ if individual } i \text{ makes choice } j \text{ at time } t, \text{ and } 0 \text{ otherwise.}$$

With all this in place, our discrete probability distribution will be defined by

$$P_{it,j} = \text{Prob}(d_{it,j} = 1 \mid \mathbf{X}_{it}, \mathbf{z}_{it}, \mathbf{u}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \dots), j = 1, \dots, J_{it},$$

where  $\mathbf{X}_{it}$  is the set of attributes of all  $J_{it}$  choices in the choice set for individual  $i$  at time  $t$ . Note that being characteristics of the individual, and not the choices,  $\mathbf{z}_{it}$  and  $\mathbf{u}_{it}$  do not vary across the choices. Whether the preference parameters,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\gamma}$ , ..., should be allowed to vary (i.e., whether they do vary) across individuals – that is, whether the parameters of the utility functions are heterogeneous – is a question that we will pursue at several points below. We will assume (not completely innocently) that in every choice situation, the individual actually makes a choice.<sup>4</sup> It follows that

$$\sum_{j=1}^{J_{it}} d_{it,j} = 1 \quad \text{and} \quad \sum_{j=1}^{J_{it}} P_{it,j} = 1. \quad (0.4)$$

The ‘model’ consists of the interesting or useful features of  $P_{it,j}$ . The preceding assumes that at time  $t$ , the consumer makes a single decision. It will be necessary in Section 0.4 to extend the model to cases of two or more decisions. This is straightforward, but requires a small change in notation and interpretation. We will defer that extension until we encounter it in the discussion in Section 0.4.

We close this section with some definitions of terms that will be used throughout the text. The individual *characteristics* such as gender or education are denoted  $\mathbf{z}_{it}$ . Attributes of the choices, such as prices, are denoted  $\mathbf{x}_{it,j}$ . We denote by *binomial* or *multinomial choice*, the single choice made between either two or more than two choices. The term *binary choice* is often used interchangeably with the former. A *bivariate choice* or *multivariate choice* is the set of 2 or more choices made in a single choice situation. In one of our applications, an individual chooses not to visit a physician or to visit at least once; this is a binomial choice. This coupled with a second

---

<sup>3</sup> Some formulations of the models, such as models of heteroscedasticity and the random parameters model, will also involve additional parameters. These will be introduced later. They are omitted at this point to avoid cluttering the notation.

<sup>4</sup> In some settings, it will be appropriate to model “none” or “the outside choice” as the  $J_{it}$ th choice.

decision, whether to visit the hospital, constitute a bivariate choice. In a different application, the choice of which of four modes to use for travel constitutes a multinomial choice.

## 0.2.2 Estimation and inference

‘Estimation’ in this setting is less clearly defined than in the familiar linear regression model. If the model is fully parametric, then the way that the parameters interact with the variables in the model, and the particular function that applies to the problem are all fully specified. The model is then

$$P_{it,j} = F(j, \mathbf{X}_{it}, \mathbf{z}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}_{it}), j = 1, \dots, J_{it}.$$

We will consider models that accommodate unobserved individual heterogeneity,  $\mathbf{u}_{it}$  in Sections 0.7 and 0.8. For the present, to avoid an inconvenience in the formulation, we consider a model involving only the observed data. Various approaches to estimation of parameters and derivative quantities in this model have been proposed, but the likelihood based estimator is by far the method of choice in the received literature. The log likelihood for the model is

$$\ln L = \sum_{i=1}^n \sum_{t=1}^{T_i} \sum_{j=1}^{J_{it}} d_{it,j} \ln F(j, \mathbf{X}_{it}, \mathbf{z}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}), i = 1, \dots, n, t = 1, \dots, T_i.$$

The maximum likelihood estimator is that function of the data that maximizes  $\ln L$ .<sup>5</sup> (See, e.g., Greene (2008, Chapter 14, for discussion of maximum likelihood estimation.) The Bayesian estimator will be the mean of the posterior density,

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{D}, \mathbf{X}, \mathbf{Z}) = \frac{L \times g(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\int_{\boldsymbol{\beta}, \boldsymbol{\gamma}} L \times g(\boldsymbol{\beta}, \boldsymbol{\gamma}) d\boldsymbol{\beta} d\boldsymbol{\gamma}}$$

where  $g(\boldsymbol{\beta}, \boldsymbol{\gamma})$  is the prior density for the model parameters and  $(\mathbf{D}, \mathbf{X}, \mathbf{Z})$  is the full sample of data on all variables in the model. (General discussions of Bayesian methods may be found in Koop (2003), Lancaster (2004) and Geweke (2005).) Semiparametric methods, generally in the index form,

$$P_{it,j} = F(j, \mathbf{X}_{ij} \boldsymbol{\beta}, \mathbf{z}_{it}' \boldsymbol{\gamma}), j = 1, \dots, J_{it}$$

but without a specific distributional assumption are common in the received literature, particularly in the analysis of binary choices and panel data. These will be considered briefly in Sections 0.3.5 and 0.7.2. Nonparametric analysis of discrete choice data is on the frontier of the theory, and does not play much of a role in the empirical literature. We will note this segment of the development briefly in Section 0.3.5 but not examine it in detail. [See Li and Racine (2007).]

Estimation and inference about model parameters is discussed in the sections to follow. Though the model is commonly formulated as an ‘index function,’ model, even in this form, it will generally bear little resemblance to the linear regression model. As in other nonlinear cases, the interpretation of the model coefficients is ambiguous. Partial effects on the probabilities associated with the choices for individual  $i$  at time  $t$  are defined as

---

<sup>5</sup> The formulation assumes that the  $T_i$  choices made by individual  $i$  are unconditionally independent. This assumption may be inappropriate. In one of our applications, the assumption is testable.

$$\delta(j, \mathbf{X}_{it}'\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) = \partial F(j, \mathbf{X}_{it}'\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) / \partial \begin{pmatrix} \mathbf{x}_{it,j} \\ \mathbf{z}_{it} \end{pmatrix} = F'(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}.$$

These are likely to be of interest for particular individuals, or averaged across individuals in the sample. A crucial implication for use of the model is that these partial effects may be quite different from the coefficients themselves. Since there is no ‘regression’ model at work, this calls into question the interpretation of the model and its parts. No generality is possible at this point. We will return to the issue below.

A related exercise in marginal analysis of the sample is to examine the aggregate outcomes predicted by the model,

$$\hat{n}_{t,j} = \sum_{i=1}^n \hat{F}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma}) = \sum_{i=1}^n \hat{d}_{it,j}$$

where the ‘^’ indicates the estimate from the model. For example, if  $x_{it,j,k}$  denotes a policy variable – a price or a tax, we might be interested in

$$\Delta \hat{n}_{t,j} = \sum_{i=1}^n \hat{F}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma} | x_{it,j,k}^1) - \sum_{i=1}^n \hat{F}(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma} | x_{it,j,k}^0).$$

Although the subject of the impact in the partial effect is already scaled – it is a probability between zero and one – it is still common for researchers to report elasticities of probabilities rather than partial effects. These are

$$\eta_{it,j}(\text{variable}_{it,j,k}) = \frac{F'(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma})}{F(j, \mathbf{X}_{it}\boldsymbol{\beta}, \mathbf{z}_{it}'\boldsymbol{\gamma})} \times \text{variable}_{it,j,k} \times \text{coefficient}_k.$$

This is prominently the case in the analysis of multinomial choice models, as we will explore in Section 0.7.

Finally, again because the model does not correspond to a regression except in a very loose sense, the concept of fit measures is also ambiguous. There is no counterpart to ‘explained variation’ or ‘total variation’ in this class of models, so the idea behind the coefficient of determination ( $R^2$ ) in linear regression has no meaning here. What is required to assess the fit of the model is first a specification of how the model will be used to predict the outcome (choice), then an assessment of how well estimated model does in that regard. Various measures are considered below.

### 0.2.3 Applications

It will be helpful in the exposition below to illustrate the computations with a few concrete examples based on ‘live’ data. We will use two familiar data sets. The RWM Health Care data (our appellation) was used in Riphahn, Wambach and Million (2003) to analyze utilization of the German health care system. The data set used is an unbalanced panel of 7,293 individual families observed from one to seven times. It is part of the German Socioeconomic Panel data set (GSOEP). These data were downloaded from the archive site of the *Journal of Applied Econometrics*. We will use these to illustrate the single equation and panel data binary and ordered choice models and models for counts presented in Sections 0.3 to 0.7. The second data set is also widely used, to illustrate multinomial choice models. These data from Hensher and Greene (e.g., 2003) are a survey of 210 travelers between Sydney and Melbourne who chose

among four modes, air, train, bus and car. We will use these data to illustrate a few multinomial choice models in Section 0.7.

### 0.3 Binary Choice

The fundamental building block in the development of discrete choice models is the basic model for choice between two alternatives. We will formulate this in a random utility framework with utility of two choices,

$$U_{i,1} = \mathbf{x}_{i,1}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,1}$$

$$U_{i,0} = \mathbf{x}_{i,0}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,0}$$

For convenience at this point, we assume there is a single choice made, so  $T_i = 1$ . The utility functions are in the index form, with characteristics and attributes and common (generic) coefficients. The random terms,  $\varepsilon_{i,1}$  and  $\varepsilon_{i,0}$  represent unmeasured influences on utility. (Looking forward, without these random terms, the model would imply that with sufficient data (and consistent parameter estimators), utility could be ‘observed’ exactly, which seems improbable at best.) Consistent with the earlier description, the analyst observes the choice most preferred by the individual, that is, the one with the greater utility, say choice 1. Thus, the observed outcome reveals that

$$U_{i,1} > U_{i,0}$$

or

$$\mathbf{x}_{i,1}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,1} > \mathbf{x}_{i,0}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,0}$$

or

$$(\mathbf{x}_{i,1}'\boldsymbol{\beta} - \mathbf{x}_{i,0}'\boldsymbol{\beta}) + (\mathbf{z}_i'\boldsymbol{\gamma} - \mathbf{z}_i'\boldsymbol{\gamma}) > (\varepsilon_{i,0} - \varepsilon_{i,1})$$

(0.15)

or

$$(\mathbf{x}_{i,1} - \mathbf{x}_{i,0})'\boldsymbol{\beta} > (\varepsilon_{i,0} - \varepsilon_{i,1}).$$

This exercise reveals several identification problems in the model as stated so far. First, we have implicitly assumed that in the event that the two utilities are equal, the consumer chooses alternative 0. This is a normalization. Recall that we assumed earlier that the individual makes exactly one choice. Second, it is evident that in describing the choice process in this fashion, it is the relative values of the attributes of the choices that matter, the difference between  $\mathbf{x}_{i,1}$  and  $\mathbf{x}_{i,0}$  is the determinant of the observed outcome, not the specific values of either. Third, note that the choice invariant component,  $\mathbf{z}_i$  has fallen out of the choice process. The implication is that unless the individual’s characteristics influence the utilities differently by  $\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_0$ , it is not possible to measure their impact on the choice process. Finally,  $\varepsilon_{i,1}$  and  $\varepsilon_{i,0}$  are random variables with so far unspecified means and variances. With respect to the means, if they are  $\mu_1$  and  $\mu_2$ , only  $\mu_0 - \mu_1$  enter the choice. As such, if the means were  $\mu_1 + \phi$  and  $\mu_0 + \phi$ , the same outcome would be observed. These means cannot be measured with observed data, so at least one, say  $\mu_0$ , is normalized to zero. Finally, consider the outcome of scaling both utilities by an arbitrary constant,  $\sigma$ . The new random components would be  $\sigma\varepsilon_{i,1} = \varepsilon_{i,1}^*$  and  $\sigma\varepsilon_{i,2} = \varepsilon_{i,2}^*$ , and  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  would be scaled likewise. However, this scaling of the model would have no impact on the observed outcome in the last line above. The same choice would be observed whatever positive value  $\sigma$  takes. Thus, there is yet one more indeterminacy in the model. This can be resolved in several ways. The most common expedient is to normalize the scaling of the random components to one.

Combining all of these, we obtain a conventional form of the model for the choice

between two alternatives,

$$\Delta U_i = \mu_1 + (\Delta \mathbf{x}_i)' \boldsymbol{\beta} + \mathbf{z}_i' (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + \varepsilon_{i0} - \varepsilon_{i1}, \quad E[\varepsilon_{i0} - \varepsilon_{i1} | \mathbf{X}_i, \mathbf{z}_i] = 0, \quad \text{Var}[\varepsilon_{i0} - \varepsilon_{i1} | \mathbf{X}_i, \mathbf{z}_i] = 1.$$

$$d_{i1} = 1 \text{ if } \Delta U_i > 0 \text{ and } d_{i1} = 0 \text{ otherwise,}$$

$$d_{i2} = 1 - d_{i1}.$$

In a more familiar arrangement that combines all these ideas, we would have

$$d_i^* = \alpha + \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i \tag{0.1}$$

$$d_i = 1 \text{ if } d_i^* > 0, \text{ and } d_i = 0 \text{ otherwise,}$$

where  $d_i = 1$  indicates choice 1 is selected.

### 0.3.1 Regression models

The preceding describes an underlying theoretical platform for a binary choice, based on a model of random utility. In order to translate it to an econometric model, we will add the assumptions behind the stochastic component of the specification,  $\varepsilon_i$ . To this point, the specification is semiparametric. We have not assumed anything specific about the underlying distribution, only that  $\varepsilon_i$  represents the random (from the point of view of the econometrician) element in the net utility function of individual  $i$ . The restrictions imposed (zero mean, unit variance) are normalizations related to the identification issue, not intended to be substantive restrictions on behavior. (Indeed, the unit variance assumption turns out to be unnecessary for some treatments as well. We will return to this below.)

We can approach the specification in (0.1) from a different viewpoint. The random utility approach specifies that  $d_i^*$  represents the strength of the individual's preference for alternative 1 relative to alternative 2. An alternative approach departs from (0.1) as a *latent regression model*. The dependent variable is assumed to be unobservable; the observation is a censored variable that measures  $d_i^*$  relative to a benchmark, zero. For an example, consider a model of loan default. One would not typically think of loan default as a utility maximizing choice. On the other hand, in the context of (0.1), one might think of  $d_i^*$  as a latent measure of the financial distress of individual  $i$ . If  $d_i^*$  is high enough, the individual defaults, we observe  $d_i = 1$ . By this construction, the appropriate model for  $d_i$  is a *censored regression*. Once we endow  $\varepsilon_i$  with a proper probability distribution, (0.1) can be construed as a regression model.

With the assumption of a specific distribution for  $\varepsilon_i$ , we obtain a statement of the choice probabilities,

$$\begin{aligned} \text{Prob}(d_i = 1 \mid \mathbf{X}_i, \mathbf{z}_i) &= \text{Prob}(d_i^* > 0 \mid \mathbf{x}_i, \mathbf{z}_i) \\ &= \text{Prob}(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i > 0). \\ &= \text{Prob}[\varepsilon_i > -(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})] \\ &= 1 - \text{Prob}[\varepsilon_i \leq -(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})] \\ &= P_{i,1}. \end{aligned}$$

It follows that

$$\begin{aligned} E[d_i | \mathbf{x}_i, \mathbf{z}_i] &= 0 \times \text{Prob}(d_i = 0 | \mathbf{x}_i, \mathbf{z}_i) + 1 \times \text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i) \\ &= \text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i) \end{aligned}$$

so we now have a regression model to manipulate as well. The implied probability endowed by our assumption of the distribution of  $\varepsilon_i$  becomes the regression of  $d_i$  on  $\mathbf{x}_i, \mathbf{z}_i$ . By this construction, one might bypass the random utility apparatus, and simply embark on modeling

$$\begin{aligned} d_i &= E[d_i | \mathbf{X}_i, \mathbf{z}_i] + a_i \\ &= \text{Prob}(d_i = 1 | \mathbf{X}_i, \mathbf{z}_i) + a_i \end{aligned}$$

where, by construction,  $a_i$  has zero mean, conditioned on the probability function. A remaining step is to construct the appropriate conditional mean function. A *linear probability model*,

$$d_i = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + a_i,$$

has been suggested in some settings. [See, e.g., Aldrich and Nelson (1984), Caudill (1988), Heckman and Snyder (1997) and Angrist (2001).] The linear probability model has some significant shortcomings, the most significant of which is that the linear function cannot be constrained to lie between zero and one, so its interpretation as a probability model is suspect. With few exceptions, including those noted, researchers have employed proper probability distributions for the implied regressions. The logit model and probit model described in the next section are the overwhelming choices in the received literature.

### 0.3.2 Estimation and inference in parametric binary choice models

A parametric model is completed by specifying a distribution for  $\varepsilon_i$ . Many candidates have been proposed, though there is little in the way of observable evidence that one can use to choose among the candidates.<sup>6</sup> For convenience, we will assume a symmetric distribution, such as the normal or logistic which are used in the overwhelming majority of studies. For a symmetric distribution,

$$\begin{aligned} 1 - \text{Prob}[\varepsilon_i \leq -(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})] &= \text{Prob}(\varepsilon_i \leq \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma}) \\ &= F(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma}). \end{aligned}$$

Once again relying on the symmetry of the distribution, the probabilities associated with the two outcomes are

$$\begin{aligned} \text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i) &= F(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma}) \\ \text{and} \\ \text{Prob}(d_i = 0 | \mathbf{x}_i, \mathbf{z}_i) &= F[-(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})]. \end{aligned}$$

For the two outcomes  $d_i = j, j = 0, 1$ , these may be combined in the form suggested earlier,

---

<sup>6</sup> See, for example, the documentation for *LIMDEP* (Econometric Software, 2007) or *Stata* (Stata, Inc., 2007).

$$F(j, \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma}) = F[(2j - 1)(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})]$$

where

$$F(c) = \Lambda(c) = \frac{\exp(c)}{1 + \exp(c)} \text{ for the logistic distribution and}$$

$$F(c) = \Phi(c) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz \text{ for the normal distribution.}$$

The assumption of the logistic distribution gives rise to the *logit model* while the normal distribution produces the *probit model*.

### Parameter estimation

The model is now fully parameterized, so the analysis can proceed based either on the likelihood function or the posterior density. We consider the maximum likelihood estimator first, and the Bayesian estimator in Section 0.3.4.

The log likelihood function for the observed data is

$$\begin{aligned} \ln L &= \sum_{i=1}^n \ln \text{Prob}(d = d_i | \mathbf{x}_i, \mathbf{z}_i) \\ &= \sum_{d_i=1}^n \ln \text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i) + \sum_{d_i=0}^n \ln \text{Prob}(d_i = 0 | \mathbf{x}_i, \mathbf{z}_i) \\ &= \sum_{i=1}^n \ln F[(2d_i - 1)(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})] \end{aligned}$$

Estimation by maximizing the log likelihood is straightforward for this model. The gradient of the log likelihood is

$$\frac{\partial \ln L}{\partial \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}} = \sum_{i=1}^n (2d_i - 1) \frac{F'[(2d_i - 1)(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})]}{F[(2d_i - 1)(\mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma})]} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} = \sum_{i=1}^n \mathbf{g}_i = \mathbf{g}.$$

The maximum likelihood estimators of the parameters are found by equating  $\mathbf{g}$  to zero, an optimization problem that requires an iterative solution.<sup>7</sup> For convenience in what follows, we will define

$$q_i = (2d_i - 1), \mathbf{w}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, c_i = q_i \mathbf{w}_i' \boldsymbol{\theta}, F_i = F(c_i), F_i' = dF_i/dc_i = f_i.$$

(Thus,  $F_i$  is the cdf and  $f_i$  is the density for the assumed distribution.) It follows that

$$\mathbf{g}_i = q_i F_i'(c_i) \mathbf{w}_i = q_i f_i \mathbf{w}_i.$$

<sup>7</sup> We are assuming that the data are 'well behaved' so that the conditions underlying the standard optimality properties of MLEs are met here. The conditions and the properties are discussed in Greene (2008). We will take them as given in what follows.

Statistical inference about the parameters is done using one of the three conventional estimators of the asymptotic covariance matrix, the Berndt et al. (1974) estimator based on the outer products of the first derivatives

$$\mathbf{V}_{\text{BHHH}} = \left[ \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i' \right]^{-1},$$

the actual Hessian,

$$\mathbf{V}_{\text{H}} = \left[ -\sum_{i=1}^n \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} = \left[ -\sum_{i=1}^n \frac{F_i F_i'' - (F_i')^2}{F_i^2} \mathbf{w}_i \mathbf{w}_i' \right]^{-1}$$

or the expected Hessian, which can be shown to equal

$$\mathbf{V}_{\text{EH}} = \left[ -\sum_{i=1}^n E_{d_i} \left( \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) \right]^{-1} = \left[ -\sum_{i=1}^n \frac{f(c_i) f(-c_i)}{F_i (1 - F_i)} \mathbf{w}_i \mathbf{w}_i' \right]^{-1}$$

[It has become common, even de rigueur, to compute a ‘robust’ covariance matrix for the MLE using  $\mathbf{V}_{\text{H}} \times \mathbf{V}_{\text{BHHH}}^{-1} \times \mathbf{V}_{\text{H}}$  under the assumption that the maximum likelihood estimator is ‘robust’ to failures of the specification of the model. In fact, there is no obvious failure of the assumptions of the model (distribution, omitted variables, heteroscedasticity, correlation across observations) for which the MLE remains consistent so the virtue of the ‘corrected’ covariance matrix is questionable. See Freedman (2006).]

For the two distributions considered here, the derivatives in the preceding are relatively simple. For the logistic,

$$\begin{aligned} F(c) &= \Lambda(c), f(c) = F'(c) = \Lambda(c)[1 - \Lambda(c)], \\ F''(c) &= F'(c)[1 - 2\Lambda(c)] = \Lambda(c)[1 - \Lambda(c)] [1 - 2\Lambda(c)]. \end{aligned}$$

For the normal distribution (probit model), the counterparts are

$$F(c) = \Phi(c), f(c) = F'(c) = \phi(c), F''(c) = -c\phi(c).$$

In both cases,  $f(c) = f(-c)$  and  $F(-c) = 1 - F(c)$ . For estimation and inference purposes, a further convenient result is, for the logistic distribution,

$$-[F(c)F''(c) - (F'(c))^2]/F(c)^2 = \Lambda(c)(1 - \Lambda(c)) > 0 \text{ for all } c$$

while for the normal distribution,

$$-[F(c)F''(c) - (F'(c))^2]/F(c)^2 = c[\phi(c)/\Phi(c)] + [\phi(c)/\Phi(c)]^2 > 0 \text{ for all } c.^8$$

The implication is that both the second derivatives matrix and the expected second derivatives matrix are negative definite for all values of the parameters and data for both models. Optimization using Newton’s method or the method of scoring will always converge to the

---

<sup>8</sup> The sign of the result for the logistic distribution is obvious. See, e.g., Maddala (1983) for a proof of the result for the normal distribution.

unique maximum of the log likelihood function, so long as the weighting matrix ( $\mathbf{V}_{\text{BHHH}}$ ,  $\mathbf{V}_H$  or  $\mathbf{V}_{\text{EH}}$ ) are not singular.<sup>9</sup>

### *Residuals and predictions*

Two additional useful results are obtained from the necessary conditions for maximizing the log likelihood function. First, the component of the score function that corresponds to the constant term is

$$\sum_{i=1}^n q_i \frac{F'(q_i \mathbf{w}'_i \boldsymbol{\theta})}{F(q_i \mathbf{w}'_i \boldsymbol{\theta})} = 0.$$

The terms in this sum are the *generalized residuals* for the model. As do the ordinary residuals in the regression model, the generalized residuals sum to zero at the MLE. These terms have been used for specification testing in this model. [See Chesher and Irish (1987).] For the logit model, it can be shown that the result above implies that

$$\frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n F(\mathbf{w}'_i \boldsymbol{\theta}) \text{ or } \frac{1}{n} \sum_{i=1}^n [d_i - F(\mathbf{w}'_i \boldsymbol{\theta})] = 0$$

when  $F$  is evaluated at the maximum likelihood estimators of the parameters. The implication is that the average of the predicted probabilities from the logit model will equal the proportion of the observations that are equal to one. A similar (albeit inexact) outcome will be seen in empirical results for the probit model. The theoretical result for the probit model has not been shown analytically.

### *Marginal effects*

Partial effects in the binary choice model are computed for continuous variables using the general result

$$\delta_i = \frac{\partial \text{Prob}(d_i = 1 | \mathbf{w}_i)}{\partial \mathbf{w}_i} = f(\mathbf{w}'_i \boldsymbol{\theta}) \boldsymbol{\theta}$$

For a binary variable such as gender or degree attained, the counterpart would be

$$\Delta_i = F(\mathbf{w}'_i \boldsymbol{\theta} + \gamma_k) - F(\mathbf{w}'_i \boldsymbol{\theta})$$

where  $\gamma_k$  is the coefficient on the dummy variable of interest (assumed to be a characteristic of the individual). These are typically evaluated for the average individual in the sample, though current practice somewhat favors the *average partial effect*,

---

<sup>9</sup> There are data configurations, in addition to simple multicollinearity, that can produce singularities. One possibility is that of a variable in  $\mathbf{x}_i$  or  $\mathbf{z}_i$  that can predict  $d_i$  perfectly based on a specific cut point in the range of that variable. Another is an empty cell in the  $2 \times 2$  cross tabulation of the binary variable  $d_i$  and a binary variable in  $\mathbf{x}_i$  or  $\mathbf{z}_i$ .

$$\begin{aligned}
\bar{\delta} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \text{Prob}(d_i = 1 | \mathbf{w}_i)}{\partial \mathbf{w}_i} \\
&= \frac{1}{n} \sum_{i=1}^n f(\mathbf{w}'_i \boldsymbol{\theta}) \boldsymbol{\theta} \\
&= \left( \frac{1}{n} \sum_{i=1}^n f(\mathbf{w}'_i \boldsymbol{\theta}) \right) \boldsymbol{\theta}.
\end{aligned}$$

(The two estimators will typically not differ substantively.) Standard errors for partial effects are usually computed using the delta method. Let  $\mathbf{V}$  denote the estimator of the asymptotic covariance matrix of the MLE of  $\boldsymbol{\theta}$ . For a particular vector,  $\mathbf{w}_i$ ,

$$\Gamma_i = \frac{\partial \delta_i}{\partial \boldsymbol{\theta}'} = [f'(\mathbf{w}'_i \boldsymbol{\theta})] \mathbf{I} + [f(\mathbf{w}'_i \boldsymbol{\theta})] \boldsymbol{\theta} \mathbf{w}'_i.$$

For a binary variable in the model in addition to (or in) the  $\mathbf{w}_i$ , the corresponding row of  $\Gamma_i$  would be

$$\Gamma_{i,k} = \partial \Delta_{i,k} / \partial (\boldsymbol{\theta}', \gamma_k) = f(\mathbf{w}'_i \boldsymbol{\theta} + \gamma_k) [\mathbf{w}_i, 1] - f(\mathbf{w}'_i \boldsymbol{\theta}) [\mathbf{w}_i, 0]$$

For the particular choice of  $\mathbf{w}_i$ , then, the estimator of the asymptotic covariance matrix for  $\delta_i$  would be  $\Gamma_i \mathbf{V} \Gamma_i'$ , computed at the maximum likelihood estimates. It is common to do this computation at the means of the data,  $\bar{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i$ . For the average partial effect, the computation is complicated a bit because the terms in  $\bar{\delta}$  are correlated – they use the same estimator of the parameters – so the variance of the mean is not  $(1/n)$  times the sum of the variances. It can be shown [see Greene (2008), chapter 23] that the appropriate computation for this computation reduces to

$$\text{Est.Asy.Var}[\bar{\delta}] = \bar{\Gamma} \mathbf{V} \bar{\Gamma}' \text{ where } \bar{\Gamma} = \frac{1}{n} \sum_{i=1}^n \Gamma_i.$$

An alternative approach to computing standard errors for the marginal effects is the method of Krinsky and Robb (1986). The a set of  $R$  random draws is taken from the estimated (asymptotic) normal population with mean  $\hat{\boldsymbol{\theta}}_{MLE}$  and variance  $\mathbf{V}$  and the empirical mean squared deviation of the estimated partial effects is computed around that computed using the MLE;

$$\text{Est.Asy.Var}[\bar{\delta}] = \frac{1}{R} \sum_{r=1}^R (\bar{\delta}_r - \bar{\delta})(\bar{\delta}_r - \bar{\delta})'$$

where  $\bar{\delta}_r$  is computed at the random draw and  $\bar{\delta}$  is computed at  $\hat{\boldsymbol{\theta}}_{MLE}$ .

An empirical conundrum can arise when doing inference about partial effects rather than coefficients. For any particular variable,  $w_k$ , the preceding theory does not guarantee that both the estimated coefficient,  $\theta_k$  and the associated partial effect,  $\delta_k$  will both be ‘statistically significant,’ or statistically insignificant. In the event of a conflict, one is left with the uncomfortable problem of simultaneously rejecting and not rejecting the hypothesis that a variable should appear in the model. Opinions differ on how to proceed. Arguably, the inference should be about  $\theta_k$ , not  $\delta_k$ , since in the latter case, one is testing a hypothesis about a function of all the coefficients, not just the one of interest.

### Hypothesis tests

Conventional hypothesis tests about restrictions on the model coefficients,  $\theta$ , can be carried out using any of the three familiar procedures. Given the simplicity of the computations for the maximum likelihood estimator, the likelihood ratio test is a natural candidate. The likelihood ratio statistic is

$$\lambda_{LR} = 2[\ln L_1 - \ln L_0]$$

where ‘1’ and ‘0’ indicate the values of the log likelihood computed at the unrestricted (alternative) estimator and the restricted (null) estimator, respectively. A hypothesis that is usually of interest in this setting is the null hypothesis that all coefficients save for the constant term are equal to zero. In this instance, it is simple to show that regardless of the assumed distribution,

$$\ln L_0 = n [P_1 \ln P_1 + P_0 \ln P_0]$$

where  $P_1$  is the proportion of observations for which  $d_i$  equals one, which is also  $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ , and  $P_0 = 1 - P_1$ .

Wald statistics use the familiar results, all based on the unrestricted model. The general procedure departs from the null hypothesis

$$H_0: \mathbf{r}(\boldsymbol{\theta}, \mathbf{c}) = \mathbf{0}$$

where  $\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})$  is a vector of  $M$  functionally independent restrictions on  $\boldsymbol{\theta}$  and  $\mathbf{c}$  is a vector of constants. The typical case is the set of linear restrictions,  $H_0: \mathbf{R}\boldsymbol{\theta} - \mathbf{c} = \mathbf{0}$ , where  $\mathbf{R}$  is a matrix of constants. The Wald statistic for testing the null hypothesis is constructed using the delta method to obtain an asymptotic covariance matrix for  $[\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})]$ . The statistic is

$$\lambda_{WALD} = [\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})]' [\mathbf{R}(\boldsymbol{\theta}, \mathbf{c}) \mathbf{V} \mathbf{R}(\boldsymbol{\theta}, \mathbf{c})]^{-1} [\mathbf{r}(\boldsymbol{\theta}, \mathbf{c})]$$

where  $\mathbf{R}(\boldsymbol{\theta}, \mathbf{c}) = \partial \mathbf{r}(\boldsymbol{\theta}, \mathbf{c}) / \partial \boldsymbol{\theta}'$  and all computations are carried out using the unrestricted maximum likelihood estimator. The standard ‘ $t$  test’ of the significance of a coefficient is the most familiar example.

The Lagrange multiplier statistic is

$$\lambda_{LM} = \mathbf{g}^{0'} \mathbf{V}^0 \mathbf{g}^0$$

where ‘0’ indicates that the computations are done using the restricted estimator and  $\mathbf{V}$  is any of the estimators of the asymptotic covariance matrix of the MLE mentioned earlier. Using  $\mathbf{V}_{BHHH}$  produces a particularly convenient computation, as well as an interesting and surprisingly simple test of the null hypothesis that all coefficients save the constant are zero. Using  $\mathbf{V}_{BHHH}$  and expanding the terms, we have

$$\lambda_{LM} = \left( \sum_{i=1}^n q_i \mathbf{w}_i f_i^0 \right)' \left( \sum_{i=1}^n q_i^2 (f_i^0)^2 \mathbf{w}_i \mathbf{w}_i' \right)^{-1} \left( \sum_{i=1}^n q_i \mathbf{w}_i f_i^0 \right).$$

An immediate simplification occurs because  $q_i^2 = 1$ . The density is computed at the restricted estimator, however obtained. If the null hypothesis is that all coefficients are zero save for the constant, then, for the logit model,  $f_i^0 = f^0 = P_1(1-P_1)$ . For the probit model, the estimator of the constant term will be  $\Phi^{-1}(P_1)$  and  $f^0 = \phi[\Phi^{-1}(P_1)]$ . Taking this constant outside the summation in  $\mathbf{g}$  leaves  $\sum_{i=1}^n q_i \mathbf{w}_i = n[P_1 \bar{\mathbf{w}}_1 - P_0 \bar{\mathbf{w}}_0]$  where  $\bar{\mathbf{w}}_1$  is the sample mean of the  $n_1$  observations with  $d_i$  equal to one and  $\bar{\mathbf{w}}_0$  is the mean of the  $n_0$  remaining observations. Note that the constant  $f^0$  falls out of the resulting statistic for both logit and probit models, and we are left with the LM statistic for testing this null hypothesis,

$$\lambda_{LM} = n^2 [P_1 \bar{\mathbf{w}}_1 - P_0 \bar{\mathbf{w}}_0]' (\mathbf{W}'\mathbf{W})^{-1} [P_1 \bar{\mathbf{w}}_1 - P_0 \bar{\mathbf{w}}_0],$$

where  $\mathbf{W}$  is the data matrix with  $i$ th row equal to  $\mathbf{w}_i'$ . As in the case of the LR statistic, the same computation is used for both the probit and logit models.

### *Specification tests*

Two specification issues are typically addressed in the context of these parametric models, heteroscedasticity and the distributional assumption. For the former, since there are no useful 'residuals' whose squares will reveal anything about scaling in the model, general approaches such as the Breusch and Pagan (1979, 1980) LM test or the White (1980) test are not available. Heteroscedasticity must be built into the model and tested parametrically. Moreover, there is no robust approach to estimation and inference that will accommodate heteroscedasticity without specifically making it part of the model. (The so called robust covariance matrix is not robust to heteroscedasticity in a binary choice setting. In linear regression, the OLS estimator and White's (1980) heteroscedasticity robust covariance matrix serve that purpose.) A common approach to modeling heteroscedasticity in parametric binary choice models is based on Harvey's (1976) exponential model,

$$d_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i, \quad E[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i] = 0, \quad \text{Var}[\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i] = [\exp(\mathbf{v}_i' \boldsymbol{\tau})]^2$$

$$d_i = 1 \text{ if } d_i^* > 0, \text{ and } d_i = 0 \text{ otherwise,}$$

where  $\mathbf{v}_i$  is a known set of variables (that does not include a constant term) and  $\boldsymbol{\tau}$  is a new parameter vector to be estimated. The adjustment of the log likelihood is fairly straightforward; the terms are changed to accommodate

$$\text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{z}_i, \mathbf{v}_i) = F[\mathbf{w}_i' \boldsymbol{\theta} / \exp(\mathbf{v}_i' \boldsymbol{\tau})].$$

Maximization of the likelihood function with respect to all the parameters is somewhat more complicated, as the function is no longer globally concave. Further complications arise in interpretation of the model. The partial effects in this augmented model are

$$\boldsymbol{\delta}_i = \frac{\partial \text{Prob}(d_i = 1 | \mathbf{w}_i, \mathbf{v}_i)}{\partial \begin{pmatrix} \mathbf{w}_i \\ \mathbf{v}_i \end{pmatrix}} = f \left( \frac{\mathbf{w}_i' \boldsymbol{\theta}}{\exp(\mathbf{v}_i' \boldsymbol{\tau})} \right) \begin{pmatrix} \boldsymbol{\theta} \\ [-(\mathbf{w}_i' \boldsymbol{\theta}) / \exp(\mathbf{v}_i' \boldsymbol{\tau})] \boldsymbol{\tau} \end{pmatrix}.$$

If  $\mathbf{w}_i$  and  $\mathbf{v}_i$  have variables in common, then the two effects are added. Whether they do or not, this once again calls into question the interpretation of the original coefficients in the model. If

$\mathbf{w}_i$  and  $\mathbf{v}_i$  do share variables, then the partial effect may have sign and magnitude that both differ from those of the coefficients,  $\boldsymbol{\theta}$ . At a minimum, as before, at least the scales of the partial effects are different from those of the coefficients.

For testing for homoscedasticity, the same three statistics as before are useable. (This is a parametric restriction on the model;  $H_0: \boldsymbol{\tau} = \mathbf{0}$ .) The derivatives of the log likelihood function are presented in Greene (2008, Chapter 23). As usual, the LM test is the simplest to carry out. The term necessary to compute the LM statistic under the null hypothesis is

$$\mathbf{g}_i = q_i f_i \left( \begin{array}{c} \mathbf{w}_i \\ (-\mathbf{w}_i' \boldsymbol{\theta}) \mathbf{v}_i \end{array} \right).$$

A second specification test of interest concerns the distribution. Silva (2001) has suggested a score (LM) test that is based on adding a constructed variable to the logit or probit model. An alternative test for testing the two competing models could be based on Vuong's (1989) statistic. Vuong's test is computed using

$$\lambda_{\text{Vuong}} = \frac{\sqrt{n} \bar{m}}{s_m} \text{ where } \bar{m} = \frac{1}{n} \sum_{i=1}^n [\ln L_i(\text{probit}) - \ln L_i(\text{logit})]$$

and  $s_m$  is the sample standard deviation. Vuong shows that under certain assumptions (likely met here for these two models),  $\lambda_{\text{Vuong}}$  has a limiting standard normal distribution. Large positive values (larger than +1.96) favor the probit model while large negative values (less than -1.96) favor the logit model. The power of these statistics for this setting remains to be investigated. As with all specification tests, the power depends crucially on the true but unknown underlying model, which may be unlike either candidate model.

### *The fit of the model*

As noted earlier, in modeling binary (or other discrete) choices, there is no direct counterpart to the  $R^2$  goodness of fit statistic. A common computation which, unfortunately in spite of its name, does not provide such a measure is the *likelihood ratio index*, which is also called the

$$\text{pseudo } R^2 = 1 - \ln L / \ln L_0$$

where  $\ln L$  is the log likelihood for the estimated model (which must include a constant term) and  $\ln L_0$  is the log likelihood function for a model that only has a constant. It is tempting to suggest that this measure measures the 'contribution' of the variables to the fit of the model. It is a statistic that is between zero and one, and it does rise unambiguously as variables are added to the model. However, the 'fit' aspect of the statistic is ambiguous, since the likelihood function is not a fit measure. As a consequence, this measure can be distressingly small in a model that contains numerous precisely measured (highly significant) coefficients. [See Wooldridge (2002) for discussion.]

This does leave open the issue of how to assess the fit of the estimated model to the data. In order to address the question, the analyst must first decide what rule will be used to predict the observed outcome using the model, then determine how successful the model (and rule) are. A natural approach, since the model predicts probabilities of events is to use the estimated probability,  $F(\mathbf{w}_i' \boldsymbol{\theta})$ . The prediction is based on a rule

$$\text{Predict } d_i = 1 \text{ if the estimated Prob}(d_i = 1 | \mathbf{w}_i) \text{ is greater than } P^* \quad (0.2)$$

where  $P^*$  is to be chosen by the analyst. The usual choice of  $P^*$  is 0.5, reasoning that if the model predicts that the event is more likely to occur than not, we should predict that it will.<sup>10</sup> A summary 2x2 table of the number of cases in which the rule predicts correctly and incorrectly can be used to assess the fit of the model. Numerous single valued functions of this tally have been suggested as counterparts to  $R^2$ . For example, Cramer (1999) proposed

$$\lambda_C = (\text{average } \hat{P}_i | d_i = 1) - (\text{average } \hat{P}_i | d_i = 0)$$

This measure counts the correct predictions, and adds a penalty for incorrect predictions. Other modifications and similar alternatives have been suggested by Efron (1978), Kay and Little (1986), Ben-Akiva and Lerman (1985) and Zavoina and McKelvey (1975).

### 0.3.3 A Bayesian estimator

The preceding has developed the classical maximum likelihood estimator for binomial choice models. A Bayesian estimator for the probit model illustrates an intriguing technique for censored data models. [This treatment builds on an example in Lancaster (2004).] The model framework is, as before,

$$d_i^* = \mathbf{w}_i' \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim N[0,1] \tag{0.3}$$

$$d_i = 1 \text{ if } d_i^* > 0, \text{ otherwise } d_i = 0. \tag{04}$$

The data consist of  $(\mathbf{d}, \mathbf{W}) = (d_i, \mathbf{w}_i), i=1, \dots, n$ . The random variable  $d_i$  has a Bernoulli distribution with probabilities

$$\text{Prob}[d_i = 1 | \mathbf{w}_i, \boldsymbol{\theta}] = \Phi(\mathbf{w}_i' \boldsymbol{\theta})$$

$$\text{Prob}[d_i = 0 | \mathbf{w}_i, \boldsymbol{\theta}] = 1 - \Phi(\mathbf{w}_i' \boldsymbol{\theta}).$$

The likelihood function for the observed data,  $\mathbf{d}$ , conditioned on  $\mathbf{W}$  and  $\boldsymbol{\theta}$  is

$$L(\mathbf{d} | \mathbf{W}, \boldsymbol{\theta}) = \prod_{i=1}^n [\Phi(\mathbf{w}_i' \boldsymbol{\theta})]^{d_i} [1 - \Phi(\mathbf{w}_i' \boldsymbol{\theta})]^{1-d_i}.$$

To obtain the posterior mean (Bayesian estimator), we assume a noninformative, flat (improper) prior for  $\boldsymbol{\theta}$ ,

$$p(\boldsymbol{\theta}) \propto 1.$$

By Bayes theorem, the posterior density would be

---

<sup>10</sup> Recall that the average predicted probability,  $\bar{\hat{P}}$  equals the average outcome in the binary choice model,  $P_1$ . To a fair approximation, the standard deviation of the predicted probabilities will equal  $[P_1(1-P_1)]^{.5}$ . If the sample is highly unbalanced, say  $P_1 < .05$  or  $P_1 > .95$ , then a predicted probability as large as (or as small as) 0.5 may become unlikely. It is common in unbalanced panels for the simple prediction rule always to predict the same value.

$$\begin{aligned}
p(\boldsymbol{\theta} | \mathbf{d}, \mathbf{W}) &= \frac{p(\mathbf{d} | \mathbf{W}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{d} | \mathbf{W}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\prod_{i=1}^n [\Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{d_i} [1 - \Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{1-d_i} (1)}{\int_{\boldsymbol{\theta}} \prod_{i=1}^n [\Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{d_i} [1 - \Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{1-d_i} (1)d\boldsymbol{\theta}}
\end{aligned}$$

and the estimator would be the posterior mean,

$$\hat{\boldsymbol{\theta}}_{\text{BAYESIAN}} = E[\boldsymbol{\theta} | \mathbf{d}, \mathbf{W}] = \frac{\int_{\boldsymbol{\theta}} \boldsymbol{\theta} \prod_{i=1}^n [\Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{d_i} [1 - \Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{1-d_i} d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} \prod_{i=1}^n [\Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{d_i} [1 - \Phi(\mathbf{w}'_i\boldsymbol{\theta})]^{1-d_i} d\boldsymbol{\theta}}.$$

Evaluation of the integrals in  $\hat{\boldsymbol{\theta}}_{\text{BAYESIAN}}$  is hopelessly complicated, but a solution using the Gibbs sampler and the technique of *data augmentation*, pioneered by Albert and Chib (1993) is surprisingly simple. We begin by treating the unobserved  $d_i$ 's in (0.3) as unknowns to be estimated, along with  $\boldsymbol{\theta}$ . Thus, the  $(K+n) \times 1$  parameter vector is  $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \mathbf{d}^*)$ . We now construct a Gibbs sampler. Consider, first,  $p(\boldsymbol{\theta} | \mathbf{d}^*, \mathbf{d}, \mathbf{W})$ . If  $d_i^*$  is known, then  $d_i$  is known. It follows that

$$p(\boldsymbol{\theta} | \mathbf{d}^*, \mathbf{d}, \mathbf{W}) = p(\boldsymbol{\theta} | \mathbf{d}^*, \mathbf{W}).$$

This posterior comes from a linear regression model with normally distributed disturbances and known  $\sigma^2 = 1$ . [See (0,3) above.] This is the standard case for Bayesian analysis of the normal linear model with an uninformative prior for the slopes and known  $\sigma^2$  – See, e.g., Koop (2003) or Greene (2008, Section 18.3.1) with the additional simplification that  $\sigma^2 = 1$ . It follows that

$$p(\boldsymbol{\theta} | \mathbf{d}^*, \mathbf{d}, \mathbf{W}) = N[\mathbf{q}^*, (\mathbf{W}'\mathbf{W})^{-1}]$$

where

$$\mathbf{q}^* = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{d}^*.$$

For  $d_i^*$ , ignoring  $d_i$  for the moment, it would follow immediately from (0.3) that

$$p(d_i^* | \boldsymbol{\theta}, \mathbf{W}) = N[\mathbf{w}'_i\boldsymbol{\theta}, 1].$$

However,  $d_i$  is informative about  $d_i^*$ . If  $d_i$  equals one, we know that  $d_i^* > 0$  and if  $d_i$  equals zero, then  $d_i^* \leq 0$ . The implication is that conditioned on  $\boldsymbol{\theta}$ ,  $\mathbf{W}$ , and  $\mathbf{d}$ ,  $d_i^*$  has a truncated (above or below zero) normal distribution. The standard notation for this is this is

$$p(d_i^* | \boldsymbol{\theta}, d_i=1, \mathbf{w}_i) = N^+[\mathbf{w}'_i\boldsymbol{\theta}, 1]$$

$$p(d_i^* | \boldsymbol{\theta}, d_i=0, \mathbf{w}_i) = N^-[\mathbf{w}'_i\boldsymbol{\theta}, 1].$$

These results set up the components for a Gibbs sampler that we can use to estimate the posterior means  $E[\boldsymbol{\theta} | \mathbf{d}, \mathbf{W}]$  and  $E[\mathbf{d}^* | \mathbf{d}, \mathbf{W}]$ .

*Gibbs sampler for the binomial probit model*

The iterations for the Gibbs sampler for the binomial probit model are computed as follows:

- (1) Compute  $\mathbf{W}'\mathbf{W}$  once at the outset and obtain  $\mathbf{L}$  such that  $\mathbf{L}\mathbf{L}' = (\mathbf{W}'\mathbf{W})^{-1}$ .
- (2) Start  $\boldsymbol{\theta}$  at any value such as  $\mathbf{0}$ .
- (3) Obtain draws  $U_{i,r}$  from the standard uniform distribution. Greene (2008, Chapter 17, shows how to transform a draw from  $U[0,1]$  to a draw from the truncated normal with underlying mean  $\mu$  and standard deviation  $\sigma$ . For this application,  $\mu = \mathbf{w}_i'\boldsymbol{\theta}$  and  $\sigma = 1$ , so the draws from  $p(\mathbf{d}^*|\boldsymbol{\theta}, \mathbf{d}, \mathbf{W})$  are obtained as

$$d_{i,r}^*(r) = \mathbf{w}_i'\boldsymbol{\theta}_{r-1} + \Phi^{-1}\left[1 - (1 - U_{i,r})\Phi(\mathbf{w}_i'\boldsymbol{\theta}_{r-1})\right] \text{ if } d_i = 1$$

$$d_{i,r}^*(r) = \mathbf{w}_i'\boldsymbol{\theta}_{r-1} + \Phi^{-1}\left[U_{i,r}\Phi(-\mathbf{w}_i'\boldsymbol{\theta}_{r-1})\right] \text{ if } d_i = 0$$

This step is used to draw the  $n$  observations on  $d_{i,r}^*(r)$

- (4) To draw an observation from the multivariate normal population of  $p(\boldsymbol{\theta} | \mathbf{d}^*, \mathbf{d}, \mathbf{W})$ , we need to draw from the normal population with mean  $\mathbf{q}_{r-1}^*$  and variance  $(\mathbf{W}'\mathbf{W})^{-1}$ . For this application, we use the results at step 3 to compute  $\mathbf{q}^* = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{d}^*(r)$ . We obtain the vector,  $\mathbf{v}$ , of  $K$  draws from the  $N[0,1]$  population, then  $\boldsymbol{\theta}(r) = \mathbf{q}^* + \mathbf{L}\mathbf{v}$ .

The iteration cycles between steps (3) and (4). This should be repeated several thousand times, discarding the burn-in draws, then the estimator of  $\boldsymbol{\theta}$  is the sample mean of the retained draws. The posterior variance is computed with the variance of the retained draws. Posterior estimates of  $d_i^*$  would typically not be useful.

This application of the Gibbs sampler demonstrates in an uncomplicated case how the algorithm can provide an alternative to actually maximizing the log likelihood. The similarity of the method to the EM algorithm [Dempster, Laird and Rubin (1977)] is not coincidental. Both procedures use an estimate of the unobserved, censored data, and both estimate  $\boldsymbol{\theta}$  by using OLS using the predicted data.

### 0.3.4 Semiparametric models

The fully parametric probit and logit models 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 literature. Most of that literature is theoretical [e.g., Matzkin (1993)]. The middle ground is occupied by a few semiparametric models that have been proposed to relax the detailed assumptions of the probit and logit specifications. The single index model of Klein and Spady (1993) has been used in several applications, including Gerfin (1996), Horowitz (1993), and Fernandez and Rodriguez-Poo (1997), and provides the theoretical platform for a number of extensions.<sup>11</sup>

The single index model departs from a regression formulation,

$$E[d_i | \mathbf{w}_i] = E[d_i | \mathbf{w}_i'\boldsymbol{\theta}].$$

Then

$$\text{Prob}(d_i = 1 | \mathbf{w}_i) = F(\mathbf{w}_i'\boldsymbol{\theta} | \mathbf{w}_i) = G(\mathbf{w}_i'\boldsymbol{\theta}),$$

where  $G$  is an unknown continuous distribution function whose range is  $[0, 1]$ . The function  $G$  is not specified a priori; it is estimated (pointwise) along with the parameters. (Since  $G$  as well as  $\boldsymbol{\theta}$  is to be estimated, a constant term is not identified; essentially,  $G$  provides the location for the index that would otherwise be provided by a constant.) The criterion function for estimation, in which subscripts  $n$  denote estimators based on the sample of  $n$  observations of their unsubscripted counterparts, is

<sup>11</sup> A symposium on semiparametric modeling is Hardle and Manski (1993).

$$\ln L_n = \frac{1}{n} \sum_{i=1}^n \{d_i \ln G_n(\mathbf{w}'_i \boldsymbol{\theta}_n) + (1 - d_i) \ln [1 - G_n(\mathbf{w}'_i \boldsymbol{\theta}_n)]\}.$$

The estimator of the probability function,  $G_n$ , is computed at each iteration using a nonparametric kernel estimator of the density at  $\mathbf{w}'_i \boldsymbol{\theta}_n$ . For the Klein and Spady estimator, the nonparametric regression estimator is

$$G_n(z_i) = \frac{\bar{d} g_n(z_i | d_i = 1)}{\bar{d} g_n(z_i | d_i = 1) + (1 - \bar{d}) g_n(z_i | d_i = 0)},$$

where  $g_n(z_i | d_i)$  is the *kernel estimate of the density* of  $z_i = \mathbf{w}'_i \boldsymbol{\theta}_n$ . This result is

$$g_n(z_i | d_i = 1) = \frac{1}{ndh_n} \sum_{m=1}^n d_j K\left(\frac{z_i - \mathbf{w}'_m \boldsymbol{\theta}_n}{h_n}\right);$$

$g_n(z_i | d_i = 0)$  is obtained by replacing  $\bar{d}$  with  $1 - \bar{d}$  in the leading scalar and  $d_m$  with  $1 - d_m$  in the summation. The scalar  $h_n$  is the bandwidth. There is no firm theory for choosing the kernel function or the bandwidth. Both Horowitz and Gerfin used the standard normal density. Two different methods for choosing the bandwidth are suggested by them. Klein and Spady provide theoretical background for computing asymptotic standard errors.

Manski's (1975, 1985, 1986, 1987) maximum score estimator is yet less parameterized than Klein and Spady's model. The estimator is based on a fitting rule,

$$\text{Maximize}_{\boldsymbol{\theta}} S_{N\alpha}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n [q_i - (1 - 2\alpha)] \text{sign}(\mathbf{w}'_i \boldsymbol{\theta}).^{12}$$

The parameter  $\alpha$  is a preset quantile, and  $q_i = 2d_i - 1$  as before. If  $\alpha$  is set to  $\frac{1}{2}$ , then the maximum score estimator chooses the  $\boldsymbol{\theta}$  to maximize the number of times that the prediction has the same sign as  $z$ . This result matches our prediction rule in (0.2) with  $P^* = 0.5$ . So for  $\alpha = 0.5$ , the maximum score estimator attempts to maximize the number of correct predictions. Since the sign of  $\mathbf{w}'\boldsymbol{\theta}$  is the same for all positive multiples of  $\boldsymbol{\theta}$ , the estimator is computed subject to the constraint that  $\boldsymbol{\theta}'\boldsymbol{\theta} = 1$ . Variants of semiparametric estimators are discussed in Li and Racine (2007), including a modification by Horowitz (1992) and an estimator suggested by Lewbel (2000).

The semiparametric estimators of  $\boldsymbol{\theta}$  are robust to variation in the distribution of the random elements in the model, and even to heteroscedasticity. Robustness is an ambiguous virtue in this context. As we have seen, the raw coefficients are of questionable value in interpreting the model – in order to translate them to useful quantities we have computed partial effects and predicted probabilities. But, the semiparametric models specifically program around the assumption of a fixed distribution. They thus sacrifice the ability to compute partial effects or probabilities. What remains is the estimator of  $\boldsymbol{\theta}$  and in some cases a covariance matrix that can be used to test significance of coefficients or test hypotheses about restrictions on structural coefficients.<sup>13</sup> Perhaps for these reasons, applied work in binary choice remains overwhelmingly dominated by the parametric models.

<sup>12</sup> See Manski (1975, 1985, 1986) and Manski and Thompson (1986). For extensions of this model, see Horowitz (1992), Charlier, Melenberg and van Soest (1995), Kyriazidou (1997) and Lee (1999)

<sup>13</sup> Bootstrapping has been used to estimate the asymptotic covariance matrix for the maximum score estimator, however, Abrevaya and Huang (2005) have recently cast doubt on the validity of that approach. No other strategy is available for statistical inference in this model.

### 0.3.5 Endogenous right hand side variables

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 instrumental variable estimator for the linear model is based on moments of the data, variances and covariances. In a 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. Generalized method of moments is a possibility. Consider the model

$$d_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \gamma z_i + \varepsilon_i$$

$$d_i = 1(d_i^* > 0)$$

$$E[\varepsilon_i | z_i] = g(z_i) \neq 0.$$

Thus,  $z_i$  is endogenous in this model. The maximum likelihood estimators considered earlier will not consistently estimate  $(\boldsymbol{\beta}, \gamma)$ . (Without an additional specification that allows us to formalize  $\text{Prob}(d_i = 1 | \mathbf{x}_i, z_i)$ , we cannot state what the MLE will, in fact, estimate.) Suppose that we have a relevant (not ‘weak’) instrumental variable,  $w_i$  such that

$$E[\varepsilon_i | w_i, \mathbf{x}_i] = 0$$

$$E[w_i z_i] \neq 0.$$

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

$$E \left[ (d_i^* - \mathbf{x}_i' \boldsymbol{\beta} - \gamma z_i) \begin{pmatrix} \mathbf{x}_i \\ w_i \end{pmatrix} \right] = \mathbf{0}.$$

However,  $d_i^*$  is not observed,  $d_i$  is, but the “residual,”  $d_i - \mathbf{x}_i' \boldsymbol{\beta} - \gamma z_i$ , would have no meaning even if the true parameters were known.<sup>14</sup> 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  $[d - \Phi(\mathbf{x}_i' \boldsymbol{\beta} + \gamma z_i)]$ , that is,

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

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.<sup>15</sup> Nothing is gained in simplicity or robustness of this approach to full information maximum likelihood estimation, which we now consider. (As

<sup>14</sup> One would proceed in precisely this fashion if the central specification were a linear probability model (LPM) to begin with. See, e.g., Eisenberg and Rowe (2006) or Angrist (2001) for an application and some analysis of this case.

<sup>15</sup> This is precisely the platform that underlies the GLIM/GEE treatment of binary choice models in, e.g. the widely used programs, SAS and *Stata* .)

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  $z_i$ . This becomes essentially a simultaneous equations model. The model equations are

$$\begin{aligned} d_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \gamma z_i + \varepsilon_i, \quad d_i = 1[d_i^* > 0], \\ z_i &= \mathbf{w}_i' \boldsymbol{\alpha} + u_i, \\ (\varepsilon_i, u_i) &\sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \sigma_u \\ \rho \sigma_u & \sigma_u^2 \end{pmatrix} \right]. \end{aligned}$$

We are assuming that there is a vector of instrumental variables,  $\mathbf{w}_i$ . Probit estimation based on  $d_i$  and  $(\mathbf{x}_i, z_i)$  will not consistently estimate  $(\boldsymbol{\beta}, \gamma)$  because of the correlation between  $z_i$  and  $\varepsilon_i$  induced by the correlation between  $u_i$  and  $\varepsilon_i$ . 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  $\text{Prob}(d_i = 1 | \mathbf{x}_i, \mathbf{w}_i) = \Phi(\mathbf{x}_i' \boldsymbol{\beta}^* + \mathbf{w}_i' \boldsymbol{\alpha}^*)$ . This will define a consistent estimator of  $\boldsymbol{\beta}^* = \boldsymbol{\beta} / (1 + \gamma^2 \sigma_u^2 + 2\gamma \sigma_u \rho)^{1/2}$  and  $\boldsymbol{\alpha}^* = \gamma \boldsymbol{\alpha} / (1 + \gamma^2 \sigma_u^2 + 2\gamma \sigma_u \rho)^{1/2}$  as the coefficients on  $\mathbf{x}_i$  and  $\mathbf{w}_i$ , respectively. (The procedure will estimate the sum of the elements of  $\boldsymbol{\beta}^*$  and  $\boldsymbol{\alpha}^*$  for any variable that appears in both  $\mathbf{x}_i$  and  $\mathbf{w}_i$ .) In addition, linear regression of  $z_i$  on  $\mathbf{w}_i$  produces estimates of  $\boldsymbol{\alpha}$  and  $\sigma_u^2$ , but there is no method of moments estimator of  $\rho$  or  $\gamma$  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  $d_i$  and  $z_i$ , which we write as the product of the conditional and the marginal densities,

$$f(d_i, z_i) = f(d_i | z_i) f(z_i).$$

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

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

where  $v_i$  is normally distributed with  $\text{Var}[v_i] = (1 - \rho^2)$ . Inserting this in the first equation, we have

$$d_i^* | z_i = \mathbf{x}_i' \boldsymbol{\beta} + \gamma z_i + (\rho / \sigma) u_i + v_i.$$

Therefore,

$$\text{Prob}[d_i = 1 | \mathbf{x}_i, u_i] = \Phi \left[ \frac{\mathbf{x}_i' \boldsymbol{\beta} + \gamma z_i + (\rho / \sigma) u_i}{\sqrt{1 - \rho^2}} \right].$$

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

$$\ln L = \sum_{i=1}^n \ln \Phi \left[ (2d_i - 1) \left( \frac{\mathbf{x}_i' \boldsymbol{\beta} + \gamma w_i + (\rho / \sigma_u)(z_i - \mathbf{w}_i' \boldsymbol{\alpha})}{\sqrt{1 - \rho^2}} \right) \right] + \ln \left[ \frac{1}{\sigma_u} \phi \left( \frac{z_i - \mathbf{w}_i' \boldsymbol{\alpha}}{\sigma_u} \right) \right]. \quad (0.5)$$

This is labeled the control function approach in the recent literature. Maximization of (0,5) is straightforward. It can be simplified further by use of the Olsen (1978) transformation,  $\tau_u = 1/\sigma_u$  and  $\pi_u = \alpha/\sigma_u$ , and by letting  $\eta = \beta / (1 - \rho^2)^{1/2}$  and  $\lambda = \gamma / (1 - \rho^2)^{1/2}$  and estimating  $\eta$ ,  $\lambda$ ,  $\tau_u$ ,  $\pi_u$  and  $\rho$ . The original parameters can be recovered from the inverse transformations, and the delta method can be used to obtain the asymptotic covariance matrix.

### 0.3.6 Panel data models

The ongoing development of large, rich panel data sets on individual and family market experiences, such as the GSOEP data we are using here, has brought attention to panel data approaches for discrete choice modeling. The extensions of familiar fixed and random effects models are not direct and bring statistical and computational issues that are not present in linear regression modeling. This section will detail the most widely used techniques. This area of research is one of the most active theoretical arenas as well. We will only have space to note the theoretical frontiers briefly in the conclusions.

#### *Panel data modeling frameworks*

The natural departure point for panel data analysis of binary choice is the extension of the fixed and random effects linear regression models. Since the models considered here are nonlinear, however, the convenient least squares and feasible generalized least squares methods are unavailable. This proves to be more than an inconvenience in this setting, as it mandates consideration of some specification issues. We will begin by considering extensions of the fixed and random effects models, then turn to more general models of individual heterogeneity, the random parameters and latent class models. The various models described here all carry over to a range of specifications. However, in the applied literature, the binary choice model is the leading case.

#### *Fixed effects model*

The fixed effects model would be

$$d_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{it}, t = 1, \dots, T_i, i = 1, \dots, n$$

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

We have made the distinction between time varying attributes and characteristics,  $\mathbf{x}_{it}$ , and time invariant characteristics,  $\mathbf{z}_i$ . The common effects,  $\alpha_i$  may be correlated with the included variables,  $\mathbf{x}_{it}$ . Since the model is nonlinear, the least squares estimator is unusable. The log likelihood is

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

In principle, direct (brute force) maximization of the function with respect to  $(\alpha_1, \dots, \alpha_n, \boldsymbol{\beta}, \boldsymbol{\gamma})$  can be used to obtain estimates of the parameters and estimates of their asymptotic standard errors. However, several issues arise.

- (1) The number of individual intercept parameters may be excessive. In our application, for example, there are 7,293 families. Direct maximization of the log likelihood function for this many parameters is likely to be difficult. This purely practical issue does have a

straightforward solution, and is, in fact, not an obstacle to estimation. See Greene (2001, 2008, Chapter 23).

- (2) As in the case of the linear model, it is not possible to estimate the parameters that apply to time invariant variables,  $\mathbf{z}_i$ . In the linear case, the transformation to group mean deviations turns these variables into columns of zeros. A similar problem arises in this nonlinear model.
- (3) For groups of observations in which the outcome variable,  $d_{it}$  is always one or always zero for  $t = 1, \dots, T_i$ , such observation groups must be dropped from the sample.
- (4) The full maximum likelihood estimator for this model is inconsistent, a consequence of the *incidental parameters problem*. [See Neyman and Scott (1948) and Lancaster (2000).] The problem arises because the number of parameters in the model,  $\alpha_i$ , rises with  $n$ . With small  $T$  or  $T_i$  this produces a bias in the estimator of  $\boldsymbol{\beta}$  that does not diminish with increase in  $n$ . The best known case, that of the logit model with  $T = 2$ , was documented by Andersen (1970), Hsiao (1986) and Abrevaya (1997), who showed analytically that with  $T = 2$ , the maximum likelihood estimator of  $\boldsymbol{\theta}$  for the binary logit model in the presence of the fixed effects will converge to  $2\boldsymbol{\theta}$ . Results for other distributions and other values of  $T$  have not been obtained analytically, and are based on Monte Carlo studies. Table 1 below, extracted from Greene (2001, 2004a,b), demonstrates the effect in the probit, logit, and ordered probit model discussed in Section 0.5. (The conditional estimator is discussed below.) The model contains a continuous variable,  $x_{it1}$  and a dummy variable,  $x_{it2}$ . The population values of both coefficients are 1.0. The results, which are consistent with other studies, e.g., Katz (2001), suggest the persistence of the “small  $T$  bias” out to fairly large  $T$ .

Table 1. Means of Empirical Sampling Distributions,  $n = 1000$  Individuals Based on 200 Replications. Table entry is  $\bar{\beta}_1, \bar{\beta}_2$ .

	$T=2$	$T=3$	$T=5$	$T=8$	$T=10$	$T=20$
	$\beta_1$ $\beta_2$	$\beta_1$ $\beta_2$	$\beta_1$ $\beta_2$	$\beta_1$ $\beta_2$	$\beta_1$ $\beta_2$	$\beta_1$ $\beta_2$
Logit	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-C <sup>a</sup>	0.994, 1.048	1.003, 0.999	0.996, 1.017	1.005, 0.988	1.002, 0.999	1.000, 1.004
Probit	2.083, 1.938	1.821, 1.777	1.589, 1.407	1.328, 1.243	1.247, 1.169	1.108, 1.068
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

<sup>a</sup>Estimates obtained using the conditional likelihood function – fixed effects not estimated.

The problems listed, particularly the last, have made the full fixed effects approach unattractive. The specification, however, remains an attractive alternative to the random effects approach considered next. Two approaches have been taken to work around the incidental parameters problem in the fixed effects model. A variety of semiparametric models have been suggested, such as Honore and Kyriazidou (2000a,b) and Honore (2002).<sup>16</sup> In a few cases, including the binomial logit model (but not the probit model), it is possible to condition the fixed effects out of the model. The operation is similar to the group mean deviations transformation in the linear regression model. For the binary logit model (omitting the time invariant variables), we have

$$\text{Prob}(d_{it} = j_{it} | \mathbf{x}_{it}) = \frac{\exp[j_{it}(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})]}{1 + \exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})} \text{ where } j_{it} \text{ is the observed value.}$$

<sup>16</sup> Much of the recent research in semiparametric and nonparametric analysis of discrete choice and limited dependent variable models has focused on how to accommodate individual heterogeneity in panel data models while avoiding the incidental parameters problem.

This is the term that enters the unconditional log likelihood function. However, conditioning on  $\sum_{t=1}^{T_i} j_{it} = S_i$ , we have the joint probability

$$\text{Prob}(d_{i1} = j_{i1}, d_{i2} = j_{i2}, \dots \mid \mathbf{x}_{it}, \sum_{t=1}^{T_i} d_{it} = S_i) = \frac{\exp(\sum_{t=1}^{T_i} j_{it} \mathbf{x}'_{it} \boldsymbol{\beta})}{\sum_{\mathbf{z}, d_{it} = S_i} \exp(\sum_{t=1}^{T_i} d_{it} \mathbf{x}'_{it} \boldsymbol{\beta})}$$

[See Rasch (1960), Andersen (1970) and Chamberlain (1980).] The denominator of the conditional probability is the summation over the different realizations of  $(d_{i1}, \dots, d_{iT_i})$  that can sum to  $S_i$ . Note that in this formulation, if  $S_i = 0$  or  $T_i$ , then there is only one way for the realizations to sum to  $S_i$ , and the one term in the denominator equals the observed result in the numerator. The probability equals one, and, as noted in point (3) above, this group falls out of the estimator. The conditional log likelihood is the sum of the logs of the joint probabilities. The log likelihood is free of the fixed effects, so the estimator has the usual properties, including consistency. This estimator was used by Cecchetti (1986) and Willis (2006) to analyze magazine price changes.

The conditional estimator is consistent, so it bypasses the incidental parameter problem. However, it does have a major shortcoming. By avoiding the estimation of the fixed effects we have precluded computation of the partial effects or estimates of the probabilities for the outcomes. So, like the robust semiparametric estimators, this approach limits the analyst to simple inference about  $\boldsymbol{\beta}$  itself. One approach that might provide some headway out of this constraint is to compute second step estimates of  $\alpha_i$ . Since we have in hand, a consistent estimator of  $\boldsymbol{\beta}$ , we treat that as known, and return to the unconditional log likelihood function. For individual  $i$ , the contribution to the log likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} \ln F[q_{it}(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})]$$

For convenience, denote the ‘known’  $\mathbf{x}'_{it} \boldsymbol{\beta}$  as  $b_{it}$ . The first order condition for maximizing  $\ln L$  with respect to  $\alpha_i$ , given the known  $\boldsymbol{\beta}$ , is

$$\partial \ln L_i / \partial b_{it} = \sum_{t=1}^{T_i} [d_{it} - F(\alpha_i + b_{it})] = 0.$$

This is one equation in one unknown that can be solved iteratively to provide an estimate of  $\alpha_i$ . The resulting estimator is inconsistent, since  $T_i$  is fixed – the resulting estimates are likely also to be highly variable because of the small sample sizes. However, the inconsistency results not because it converges to something other than  $\alpha_i$ . The estimator is inconsistent because its variance is  $O(1/T_i)$ . As such, an estimator of the average partial effects,

$$\bar{\delta} = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^{T_i} f(\hat{\alpha}_i + \mathbf{x}'_{it} \hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\beta}}$$

may yet provide a useful estimate of the partial effects. This estimator remains to be examined empirically or theoretically.

The fixed effects model has the attractive aspect that it is a robust specification. The four shortcomings listed above, especially items (2) and (4) do reduce its appeal, however. The wisdom from the linear model does not carry over to binary choice models because the estimation and inference problem change substantively in nonlinear settings. The statistical aspects of the random effects model discussed next are more appealing. However, the model assumption of orthogonality of the unobserved heterogeneity and the included variables is also unattractive. The

Mundlak (1978) device is an intermediate step between these two that is sometimes used. The approach relies on a projection of the effects on the time invariant characteristics and group means of the time variables;

$$\alpha_i = \mathbf{z}_i'\boldsymbol{\gamma} + \pi_0 + \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \sigma_u u_i \text{ where } E[u_i | \bar{\mathbf{x}}_i] = 0 \text{ and } \text{Var}[u_i | \bar{\mathbf{x}}_i] = 1.$$

(The location parameter  $\pi_0$  accommodates a nonzero mean while the scale parameter,  $\sigma_u$ , picks up the variance of the effect. So the assumptions of zero mean and unit variance for  $u_i$  are just normalizations.) Inserting this result in the fixed effects model produces a type of random effects model,

$$d_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \pi_0 + \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \sigma_u u_i + \varepsilon_{it}, \quad t = 1, \dots, T_i, \quad i = 1, \dots, n$$

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

If the presence of the projection on the group means successfully picks up the correlation between  $\alpha_i$  and  $\mathbf{x}_{it}$ , then the parameters  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \pi_0, \boldsymbol{\pi}, \sigma_u)$  can be estimated by maximum likelihood as a random effects model. The remaining assumptions (functional form, distribution) are assumed to hold (at least approximately), so that the random effects treatment is appropriate.

#### *Random Effects Models and Estimation*

As suggested in the preceding, section, the counterpart to a random effects model for binary choice would be

$$d_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \sigma_u u_i + \varepsilon_{it}, \quad t = 1, \dots, T_i, \quad i = 1, \dots, n,$$

where  $E[u_i | \mathbf{x}_{it}] = 0$  and  $\text{Var}[u_i | \mathbf{x}_{it}] = 1$  and

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

(Since the random effects model can accommodate time invariant characteristics, we have reintroduced  $\mathbf{z}_i$  in the model.) The random effects model is fit by maximum likelihood assuming normality for  $\varepsilon_{it}$  and  $u_i$ . (The most common application is the random effects probit model.)

To begin, suppose the common effect is ignored, and the 'pooled' model is fit by simple ML, ignoring the presence of the heterogeneity. The (incorrectly) assumed model is

$$\text{Prob}(d_{it} = 1 | \mathbf{x}_{it}) = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}).$$

In the presence of  $u_i$ , the correct model is

$$\begin{aligned} \text{Prob}(d_{it} = 1 | \mathbf{x}_{it}) &= \text{Prob}(\varepsilon_{it} + \sigma_u u_i < \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) \\ &= \text{Prob}\left(\frac{\varepsilon_{it} + \sigma_u u_i}{\sqrt{1 + \sigma_u^2}} < \frac{\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}}{\sqrt{1 + \sigma_u^2}}\right) \\ &= \text{Prob}(v_{it} < \mathbf{x}_{it}'\boldsymbol{\beta}^u + \mathbf{z}_i'\boldsymbol{\gamma}^u), \quad v_{it} \sim N[0, 1]. \end{aligned}$$

Thus, the marginal probability that  $d_{it}$  equals one obeys the assumptions of the familiar probit. However, the coefficient vector is not  $\boldsymbol{\beta}$ , it is  $\boldsymbol{\beta}^u = \boldsymbol{\beta}/(1 + \sigma_u^2)^{1/2}$  and likewise for  $\boldsymbol{\gamma}$ . The upshot is

that ignoring the heterogeneity (random effect) is not so benign here as in the linear regression model. In the regression case, ignoring a random effect that is uncorrelated with the included variables produces an inefficient, but consistent estimator.

In spite of the preceding result, it has become common in the applied literature to report ‘robust,’ ‘cluster corrected’ asymptotic covariance matrices for pooled estimators such as the MLE above. The underlying justification is that while the MLE may be consistent (though it rarely is, as exemplified above), the asymptotic covariance matrix should account for the correlation across observations within a group. The corrected estimator is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\theta}}_{MLE}] = \left[ \sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{H}_{it} \right]^{-1} \left[ \sum_{i=1}^n \left( \sum_{t=1}^{T_i} \mathbf{g}_{it} \right) \left( \sum_{t=1}^{T_i} \mathbf{g}'_{it} \right) \right] \left[ \sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{H}_{it} \right]^{-1}$$

where  $\mathbf{H}_{it} = \partial^2 \ln F(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma})) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$  and  $\mathbf{g}_{it} = \partial \ln F(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma})) / \partial \boldsymbol{\theta}$  and all terms are computed at the pooled MLE. The estimator has a passing resemblance to the White (1980) estimator for the least squares coefficient estimator. However, the usefulness of this estimator rests on the assumption that the pooled estimator is consistent, which will generally not be the case.

Efficiency is a moot point for this estimator, since the probit MLE estimates  $\boldsymbol{\beta}$  with a bias toward zero;

$$\begin{aligned} \text{plim } \hat{\boldsymbol{\beta}}_{MLE} &= \boldsymbol{\beta}^u \\ &= \boldsymbol{\beta} / (1 + \sigma_u^2)^{1/2} \\ &= \boldsymbol{\beta} (1 - \rho^2)^{1/2}, \end{aligned}$$

where  $\rho^2 = \text{Corr}^2[\varepsilon_{it} + u_i, \varepsilon_{is} + u_i]$  for  $t \neq s$ . Wooldridge (2002) suggests that this may not be an issue here, since the real interest is in the partial effects, which are, for the correct model,

$$\boldsymbol{\delta}_{it} = \partial \text{Prob}[d_{it} = 1 \mid \mathbf{x}_{it}, \mathbf{z}_i] \partial \mathbf{x}_{it} = \boldsymbol{\beta}^u \phi(\mathbf{x}_{it}'\boldsymbol{\beta}^u + \mathbf{z}'_i\boldsymbol{\gamma}^u).$$

These would then be averaged over the individuals in the sample. It follows, then, that the ‘pooled’ estimator, that ignores the heterogeneity does not estimate the structural parameters of the model correctly, but it does produce an appropriate estimator of the average partial effects.

In the random effects model, the observations are not statistically independent – because of the common  $u_i$ , the observations  $(d_{i1}, \dots, d_{iT_i}, u_i)$  constitute a  $T_i+1$  variate random vector. The contribution of observation  $i$  to the log likelihood is this joint density, which we write

$$f(d_{i1}, \dots, d_{iT_i}, u_i \mid \mathbf{X}_i) = f(d_{i1}, \dots, d_{iT_i} \mid \mathbf{X}_i, \mathbf{z}_i, u_i) f(u_i).$$

Conditioned on  $u_i$ , the  $T_i$  random outcomes,  $d_{i1}, \dots, d_{iT_i}$ , are independent. This implies that (with the normality assumption now incorporated in the model) the contribution to the log likelihood is

$$\ln L_i = \ln \left\{ \left[ \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right] \phi(u_i) \right\},$$

where  $\phi(u_i)$  is the standard normal density. This joint density contains the unobserved  $u_i$ , which must be integrated out of the function to obtain the appropriate log likelihood function in terms of the observed data. Combining all terms, we have the log likelihood for the observed sample,

$$\ln L = \sum_{i=1}^n \ln \left[ \int_{-\infty}^{\infty} \left( \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right) \phi(u_i) du_i \right] \quad (0.6)$$

Maximization of the log likelihood with respect to  $(\boldsymbol{\beta}, \sigma_u)$  requires evaluation of the integrals in (0.6). Since these do not exist in closed form, some method of approximation must be used. The most common approach is the Hermite quadrature method suggested by Butler and Moffitt (1982). The approximation is written

$$\int_{-\infty}^{\infty} \left( \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right) \phi(u_i) du_i \approx \frac{1}{\sqrt{\pi}} \sum_{h=1}^H w_h \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sqrt{2}\sigma_u z_h))$$

where  $w_h$  and  $z_h$  are the weights and nodes of the quadrature [See Abramovitz and Stegun (1971)] and  $H$  is the number of nodes chosen (typically 20, 32 or 64). An alternative approach to the approximation is suggested by noting that

$$\left[ \int_{-\infty}^{\infty} \left( \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right) \phi(u_i) du_i \right] = E_{u_i} \left[ \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right].$$

The expected value can be approximated satisfactorily by simulation by using a sufficiently large sample of random draws from the population of  $u_i$ ;

$$\left[ \int_{-\infty}^{\infty} \left( \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_i)) \right) \phi(u_i) du_i \right] \approx \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_{ir})).$$

Sampling from the standard normal population is straightforward using modern software [see Greene (2008, Chapter 17)]. The right hand side converges to the left hand side as  $R$  increases [so long as  $\sqrt{n}/R \rightarrow 0$  – see Gourieroux and Monfort (1996)].<sup>17</sup> The *simulated log likelihood* to be maximized is

$$\ln L_S = \sum_{i=1}^n \ln \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \sigma_u u_{ir})).$$

Recent research in numerical methods has revealed alternative approaches to random sampling to speed up the rate of convergence in the integration. Halton sequences [see Bhat (1999), for example] are often used to produce approximations which provide comparable accuracy with far fewer draws than the simulation approach.

### *Dynamic Models*

An important extension of the panel data treatment in the previous section is the dynamic model,

$$d_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \lambda d_{i,t-1} + \alpha_i + \varepsilon_{it} \quad (0.7)$$

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

<sup>17</sup> The requirement does not state how large  $R$  must be, only that it ‘increase’ faster than  $n^{1/2}$ . In practice, analysts typically use several hundred, perhaps up to 1,000 random draws for their simulations.

Recent applications include Hyslop's (1999) analysis of labor force participation, Wooldridge's (2005) study of union membership and Contoyannis et al's (2004) analysis of self reported health status in the British Household Panel Survey.<sup>18</sup> In these and other applications, the central feature is *state dependence*, or the *initial conditions problem*. The individual tends to 'stick' with their previous position. Wooldridge (2002) lays out conditions under which an appropriate treatment is to model the individual effect as being determined by the initial value in

$$\alpha_i = \alpha_0 + \alpha_1 d_{i0} + \bar{\mathbf{x}}_i' \boldsymbol{\pi} + \sigma_u u_i, u_i \sim N[0,1]. \quad (0.8)$$

This is the Mundlak treatment suggested earlier with the addition of the initial state in the projection.<sup>19</sup> Inserting (0.7) in (0.8) produces an augmented random effects model that can be estimated, as in the static case, by Hermite quadrature or maximum simulated likelihood.

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000a,b) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two period panel with fixed effects. An extension of lagged effects to a parametric model is Chamberlain (1980), Jones and Landwehr (1988) and Magnac (1997) who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits its usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In this framework, the narrow assumptions about the independent variables once again limit its practical applicability. Honore and Kyriazidou (2001) have combined the logic of the conditional logit model and Manski's maximum score estimator. They specify

$$\begin{aligned} \text{Prob}(d_{i0} = 1 | \mathbf{X}_i, \mathbf{Z}_i, \alpha_i) &= F_0(\mathbf{X}_i, \mathbf{Z}_i, \alpha_i) \text{ where } \mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}), \\ \text{Prob}(d_{it} = 1 | \mathbf{X}_i, \mathbf{Z}_i, \alpha_i, d_{i0}, d_{i1}, \dots, d_{i,t-1}) &= F(\mathbf{x}_{it}' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \alpha_i + \lambda d_{i,t-1}) \quad t = 1, \dots, T \end{aligned}$$

The analysis assumes a single regressor and focuses on the case of  $T = 3$ . The resulting estimator resembles Chamberlain's but relies on observations for which  $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$  which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of  $x_{it}$  is a considerable restriction, and the authors propose a kernel density estimator for the difference,  $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$ , instead which does relax that restriction a bit. The end result is an estimator which converges (they conjecture) but to a nonnormal distribution and at a rate slower than  $n^{-1/3}$ .

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b, 1981c), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993) and Beck et al. (2001) to name a few.<sup>20</sup>

<sup>18</sup> See, as well, Hsiao (2003) for a survey of dynamic panel data models and other applications by van Doorslaer (1987), Wagstaff (1993) and Vella and Verbeek (1999).

<sup>19</sup> This is the formulation used by Contoyannis et al. Wooldridge suggested, instead, that the projection be upon all of the data,  $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots)$ . Two major practical problems with this approach are that in a model with a large number of regressors, which is common when using large, elaborate panel data sets, the number of variables in the resulting model will become excessive. Second, this approach breaks down if the panel is unbalanced, as it was in the Contoyannis et al. study.

<sup>20</sup> Beck et al. (2001) is a bit different from the others mentioned in that in their study of "state failure," they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to  $T$  appropriate. They

*Parameter Heterogeneity: Random Parameters and Latent Class Models*

Among the central features of panel data treatments of individual data is the opportunity to model individual heterogeneity, both observed and unobserved. The preceding discussion develops a set of models in which latent heterogeneity is embodied in the additive effect,  $\alpha_i$ . We can extend the model to allow heterogeneity in the other model parameters as well,. The resulting specification is

$$d_{it}^* = \mathbf{w}_{it}'\boldsymbol{\theta}_i + \alpha_i + \varepsilon_{it}, d_{it} = 1(d_{it}^* > 0).$$

The specification is completed by the assumptions about the process that generates the individual specific parameters. Note that in this formulation, the ‘effect,’  $\alpha_i$  is now merely an individual specific constant term. It is thus convenient to absorb it into the rest of the parameter vector,  $\boldsymbol{\theta}_i$  and assume that  $\mathbf{w}_{it}$  contains a constant.

A random parameters model (or mixed model or hierarchical model) in which parameters are continuously distributed across individuals, can be written

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \Delta\mathbf{z}_i + \Gamma\mathbf{u}_i$$

where  $\mathbf{u}_i$  is a set of uncorrelated random variables with means zero (means are absorbed in  $\boldsymbol{\theta}_0$ ) and variances 1 (nonunit variances are contained in the parameter matrix  $\Gamma$ ). The random effects model examined earlier emerges if  $\Delta = \mathbf{0}$  and the only random component in  $\boldsymbol{\theta}_i$  is the constant term, in which case,  $\Gamma$  would have a single nonzero diagonal element equal to  $\sigma_u$ . For the more general case, we have a random parameters formulation in which

$$E[\boldsymbol{\theta}_i | \mathbf{z}_i] = \boldsymbol{\theta}_0 + \Delta\mathbf{z}_i$$

$$\text{Var}[\boldsymbol{\theta}_i | \mathbf{z}_i] = \Gamma\Gamma'$$

A random parameters model of this sort can be estimated by Hermite quadrature [See Rabe-Hesketh et al. (2005) or by maximum simulated likelihood. [See Train (1999) and Greene (2008, Chapters 17 and 23).] The simulated log likelihood function for this model will be

$$\ln L_S = \sum_{i=1}^n \ln \frac{1}{R} \sum_{r=1}^R \prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{w}'_{it}(\boldsymbol{\theta}_0 + \Delta\mathbf{z}_i + \Gamma\mathbf{u}_{ir})))$$

Partial effects in this model can be computed by averaging the partial effects at the population conditional means of the parameters,  $E[\boldsymbol{\theta}_i | \mathbf{z}_i] = \boldsymbol{\theta}_0 + \Delta\mathbf{z}_i$ .

### 0.3.7 Application

In "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" by Riphahn, Wambach and Million (2003), the authors were interested in counts of physician visits and hospital visits. In this application, they were particularly interested in the impact that the presence of private insurance had on the utilization counts of interest, i.e., whether the data contain evidence of moral hazard. The raw data are published and available for download on the *Journal of Applied Econometrics* data archive website, The URL is given below. The sample is

---

can analyze the data essentially in a time series framework. Sepanski (2000) is another application which combines state dependence and the random coefficient specification of Akin, Guilkey and Sickles (1979).

an unbalanced panel of 7,293 households. The number of observations varies from one to seven (1,525, 1,079, 825, 926, 1,051, 1000, 887) with a total number of observations of 27,326. The variables in the data file are listed in Table 2. (Only a few of these were used in our applications.)

**Table 2. Variables in German Health Care Data File**

Variable		Mean	Standard Deviation
YEAR	calendar year of the observation	1987.82	3.17087
AGE	age in years	43.5257	11.3302
FEMALE	female = 1; male = 0	.478775	.499558
MARRIED	married = 1; else = 0	.758618	.427929
HKIDS	children under age 16 in the household = 1; else = 0	.402730	.490456
HHNINC	household nominal monthly net income in German marks / 10000	.352084	.176908
WORKING	employed = 1; else = 0	.677048	.467613
BLUEC	blue collar employee = 1; else = 0	.243761	.429358
WHITEC	white collar employee = 1; else = 0	.299605	.458093
SELF	self employed = 1; else = 0	.0621752	.241478
BEAMT	civil servant = 1; else = 0	.0746908	.262897
EDUC	years of schooling	11.3206	2.32489
HAUPTS	highest schooling degree is Hauptschul = 1; else = 0	.624277	.484318
REALS	highest schooling degree is Realschul = 1; else = 0	.196809	.397594
FACHHS	highest schooling degree is Polytechnical = 1; else = 0	.0408402	.197924
ABITUR	highest schooling degree is Abitur = 1; else = 0	.117031	.321464
UNIV	highest schooling degree is university = 1; else = 0	.0719461	.258403
HSAT	health satisfaction, 0 - 10	6.78543	2.29372
NEWHSAT***	health satisfaction, 0 - 10	6.78566	2.29373
HANDDUM	handicapped = 1; else = 0	.214015	.410028
HANDPER	degree of handicap in pct, 0 - 100	7.01229	19.2646
DOCVIS	number of doctor visits in last three months	3.18352	5.68969
DOCTOR**	1 if DOCVIS > 0, 0 else	.629108	.483052
HOSPVIS	number of hospital visits in last calendar year	.138257	.884339
HOSPITAL**	1 of HOSPVIS > 0, 0 else	.0876455	.282784
PUBLIC	insured in public health insurance = 1; else = 0	.885713	.318165
ADDON	insured by add-on insurance = 1; else = 0	.0188099	.135856

Data source: <http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/>. From Riphahn, R., A. Wambach and A. Million "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, 18, 4, 2003, pp. 387-405.

Notes: \* NEWHSAT = HSAT; 40 observations on HSAT recorded between 6 and 7 were changed to 7.

\*\* Transformed variable not in raw data file.

The model to be examined here (not the specification used in the original study) is

$$\text{Prob}(\text{Doctor}_{it} = 1 | \mathbf{x}_{it}) = F(\beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Income}_{it} + \beta_4 \text{Kids}_{it} + \beta_5 \text{Education}_{it} + \beta_6 \text{Married}_{it}).$$

(In order to examine fixed effects models, we have not used any of the time invariant variables, such as gender.) Table 3 lists the maximum likelihood estimates and estimated asymptotic standard errors for several model specifications. Estimates of the logit model are shown first, followed by the probit estimates. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. The empirical regularity that the MLE's of the coefficients in the logit model are typically about 1.6 times their probit counterparts is strikingly evident in these results (e.g., the ratios are 1.613 and 1.597 for the coefficients on age and income, respectively). The apparent differences between the logit and probit results are resolved by a comparison of the partial effects shown in Table 3. As anticipated, the results are essentially the same for the two models. The first two rows of partial effects in Table 3 compare the partial effects computed at the means of the variables in the first row to the average partial effects, computed by averaging the individual partial effects, in the second. As might be expected, the difference between them is inconsequential.

The log likelihood for the probit model is slightly larger than for the logit, however, it is not possible to compare the two on this basis – the models are not nested. The Vuong statistic based on  $v_i = \ln L_i(\text{logit}) - \ln L_i(\text{probit})$  equals -7.44, which favors the probit model. The aggregated prediction of the pooled logit model is shown below. The usual prediction rule in (0.2),  $P^* = .5$  produces the following results:

	Predicted	
Actual	0	1
0	378	9757
1	394	16797

Thus, we obtain correct prediction of  $(378+16797)/27326 = 62.9\%$  of the observations. In spite of this apparently good model performance, the pseudo- $R^2$  is only  $1 - (-17673.10) / (-18019.55) = 0.01923$ . This suggests the disconnection between these two measures of model performance. As a final check on the model itself, we tested the null hypothesis that the five coefficients other than the constant term are zero in the probit specification. The likelihood ratio test is based on the statistic

$$\lambda_{LR} = 2[-17670.94 - 27326(.37089 \ln .37089 + .62911 \ln .62911)] = 697.22.$$

The Wald statistic based on the full model is  $\lambda_{WALD} = 686.991$ . The LM statistic is computed as

$$\lambda_{LM} = \mathbf{g}_0' \mathbf{X} (\mathbf{G}_0' \mathbf{G}_0)^{-1} \mathbf{X}' \mathbf{g}_0$$

where  $\mathbf{g}_0$  is the derivative of the full log likelihood when the estimated model contains only a constant term. This is equal to  $q_{it} \phi(q_{it} \beta_0) / \Phi(q_{it} \beta_0)$  where  $\beta_0 = \Phi^{-1}(.62911) = .32949$ . Then the  $i$ th row of  $\mathbf{G}$  is  $g_{it,0}$  times the corresponding row of  $\mathbf{X}$ . The value of the LM statistic is 715.97. The 95% critical value from the chi squared distribution with 5 degrees of freedom is 11.07, so in all three cases, the null hypothesis that the slopes are zero is rejected.

The second set of probit estimates were computed using the Gibbs sampler and a noninformative prior. We used only 500 replications, and discarded the first 100 for the burn in. The similarity to the maximum likelihood estimates is what one would expect given the large sample size. We note however, that, notwithstanding the striking similarity of the Gibbs sampler to the MLE, this is not an efficient method of estimating the parameters of a probit model. The

estimator requires generation of thousands of samples of potentially thousands of observations. We used only 500 replications to produce the results in Table 3. The computations took about five minutes. Using Newton's method to maximize the log likelihood directly took less than five seconds. Unless one is wedded to the Bayesian paradigm, on strictly practical grounds, the MLE would be the preferred estimator.

Table 3 also lists the probit and logit random and fixed effects estimates. The random effects estimators produce a reasonably large estimate of  $\rho^2$ , roughly 0.44. The high correlation across observations does cast some doubt on the validity of the pooled estimator. The pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line marked "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimators, it is necessary to drop from the sample any groups that have  $Doctor_{it}$  equal to zero or one for every period. There were 3,046 such groups, which is about 42% of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient,  $\rho^2$ , is computed as  $\sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$ . For the probit model,  $\sigma_\varepsilon^2 = 1$ . The MSL estimator computes  $s_u = 0.9088376$ , from which we obtained  $\rho^2$ . The estimated partial effects for the models are also shown in Table 3. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

Finally, we carried out two tests of the stability of the model. All of the estimators listed in Table 3 derive from a model in which it is assumed that the same coefficient vector applies in every period. To examine this assumption, we carried out a homogeneity test of the hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_T$$

for the seven periods in the sample. The likelihood ratio statistic is

$$\lambda = 2 \left[ \left( \sum_{t=1}^T \ln L_t \right) - \ln L_{POOLED} \right]$$

The first part is obtained by dividing the sample into the seven years of data – the number of observations varies (3,874, 3,794, 3,792, 3,661, 4,483, 4,340, 3,377) – then estimating the model separately for each year. The calculated statistic is 202.97. The critical value from the chi squared distribution with  $(T-1)6 = 36$  degrees of freedom is 50.998, so the homogeneity assumption is rejected by the data. As a second test, we separated the sample into men and women and once again tested for homogeneity. The likelihood ratio test statistic is

$$\begin{aligned} \lambda &= 2[\ln L_{FEMALE} + \ln L_{MALE} - \ln L_{POOLED}] \\ &= 2[(-7855.219377) + (-9541.065897) - (-18019.55)] \\ &= 1246.529452. \end{aligned}$$

The critical value from the chi squared distribution with 6 degrees of freedom is 12.592, so this hypothesis is rejected as well.

**Table 3 Estimated Parameters for Panel Data Binary Choice Models**

Model	Estimate	Variable						
		<i>ln L</i>	<i>Constant</i>	<i>Age</i>	<i>Income</i>	<i>Kids</i>	<i>Education</i>	<i>Married</i>
Logit Pooled	$\beta$ (ME) [APE]	-17673.10	<b>0.25112</b>	<b>0.020709</b> (0.00481) [0.00471]	<b>-0.18592</b> (-0.0432) [-0.0423]	<b>-0.22947</b> (-0.0536) [-0.0522]	<b>-0.045587</b> (-0.0106) [-0.0104]	<b>0.085293</b> (0.0199) [0.0194]
	St.Er.		0.091135	0.001285	0.075064	0.029537	0.005646	0.033286
	Rob.SE <sup>e</sup>		0.12827	0.001743	0.091546	0.038313	0.008075	0.045314
Logit R.E. $\rho^2=0.41607$	$\beta$ (ME)	-15261.90	<b>-0.13460</b>	<b>0.039267</b> (0.00642)	<b>0.021914</b> (0.00358)	<b>-0.21598</b> (-0.0354)	<b>-0.063578</b> (-0.0103)	<b>0.025071</b> (0.00410)
	St.Er.		0.17764	0.002465	0.11866	0.047738	0.011322	0.056282
Logit F.E.(U) <sup>a</sup>	$\beta$ (ME)	-9458.64		<b>0.10475</b> (0.0249)	<b>-0.060973</b> (-0.0145)	<b>-0.088407</b> (-0.0210)	<b>-0.11671</b> (-0.0277)	<b>-0.057318</b> (-0.0136)
	St.Er.		0.007255	0.17829	0.074399	0.066749	0.10609	
Logit F.E.(C) <sup>b</sup>	$\beta$ (ME)	-6299.02		<b>0.084760</b> (0.00730)	<b>-0.050383</b> (-0.00434)	<b>-0.077764</b> (-0.00670)	<b>-0.090816</b> (-0.00782)	<b>-0.052072</b> (-0.00448)
	St.Er.		0.006502	0.15888	0.066282	0.056673	0.093044	
Probit Pooled	$\beta$ (ME)	-17670.94	<b>0.15500</b>	<b>0.012835</b> (0.00484)	<b>-0.11643</b> (-0.0439)	<b>-0.14118</b> (-0.0534)	<b>-0.028115</b> (-0.0106)	<b>0.052260</b> (0.0198)
	St.Er.		0.056516	0.000790	0.046329	0.018218	0.003503	0.020462
	Rob.SE <sup>e</sup>		0.079591	0.001074	0.056543	0.023614	0.005014	0.027904
Bayesian Pooled Probit	$\beta$ (Mean)	N/A	<b>0.15729</b>	<b>0.012807</b>	<b>-0.11319</b>	<b>-0.14160</b>	<b>-0.028234</b>	<b>0.050943</b>
	$\beta$ (Var.)		0.057824	0.000784	0.048868	0.017385	0.003437	0.020729
Probit:RE <sup>c</sup> $\rho^2=0.44789$	$\beta$ (ME)	-16273.96	<b>0.034113</b>	<b>0.020143</b> (0.00560)	<b>-0.003176</b> (-0.00088)	<b>-0.15379</b> (-0.0428)	<b>-0.033694</b> (-0.00938)	<b>0.016325</b> (0.00454)
	St.Er.		0.096354	0.001319	0.066672	0.027043	0.006289	0.031347
Probit:RE <sup>d</sup> $\rho^2=0.44799$	$\beta$ (ME)	-16279.97	<b>0.033290</b>	<b>0.020078</b> (0.00715)	<b>-0.002973</b> (-0.00106)	<b>-0.153579</b> (-0.0547)	<b>-0.033489</b> (-0.0119)	<b>0.016826</b> (0.00599)
	St.Er.		0.063229	0.000901	0.052012	0.020286	0.003931	0.022771
Probit F.E.(U)	$\beta$ (ME)	-9453.71		<b>0.062528</b> (0.0239)	<b>-0.034328</b> (-0.0132)	<b>-0.048270</b> (-0.0185)	<b>-0.072189</b> (-0.0277)	<b>-0.032774</b> (-0.0126)
	St.Er.		0.004322	0.10745	0.044559	0.040731	0.063627	

<sup>a</sup> Unconditional fixed effects estimator,

<sup>b</sup> Conditional fixed effects estimator,

<sup>c</sup> Butler and Moffitt Estimator

<sup>d</sup> Maximum simulated likelihood estimator

<sup>e</sup> Robust, "cluster" corrected standard error

## 0.4 Bivariate and multivariate binary choice

The health care data contain two binary variables, *Doctor* and *Hospital*, that one would expect to be at least correlated if not jointly determined. The extension of the binary choice to more than one choice is relatively uncomplicated, but does bring new statistical issues as well as new practical complications. We consider several two equation specifications first, as these are the leading cases, then consider the extension to an arbitrary number of binary choices.

### 0.4.1 Bivariate binary choice

A two equation binary choice model would take the form of a seemingly unrelated regressions model,

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, \end{aligned}$$

where ‘1’ and ‘2’ distinguish the equations (and are distinct from the periods in a panel data case). The bivariate binary choice model arise when the two disturbances are correlated. There is no convenient approach for this model based on the logistic model, so we assume bivariate normality at the outset. The bivariate probit model has

$$F(\varepsilon_{i,1}, \varepsilon_{i,2}) = N_2[(0,0), (1,1), \rho], \quad -1 < \rho < 1.$$

The probability associated with the joint event  $d_{i,1} = d_{i,2} = 1$  is then

$$\text{Prob}(d_{i,1} = 1, d_{i,2} = 1 \mid \mathbf{w}_{i,1}, \mathbf{w}_{i,2}) = \Phi_2[\mathbf{w}'_{i,1}\boldsymbol{\theta}_1, \mathbf{w}'_{i,2}\boldsymbol{\theta}_2, \rho]$$

where  $\Phi_2[c_1, c_2, \rho]$  denotes the bivariate normal cdf. The log likelihood function is the joint density for the observed outcomes. By extending the formulation of the univariate probit model in the preceding section, we obtain

$$\ln L = \sum_{i=1}^n \ln \Phi_2\left[\left(q_{i,1}\mathbf{w}'_{i,1}\boldsymbol{\theta}_1\right), \left(q_{i,2}\mathbf{w}'_{i,2}\boldsymbol{\theta}_2\right), \left(q_{i,1}q_{i,2}\rho\right)\right]$$

The bivariate normal integral does not exist in closed form, and must be approximated, typically with Hermite quadrature.

The model is otherwise conventional and the standard conditions for maximum likelihood estimators are obtained. Interpretation of the model does bring some complications, however. First,  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  are not the slopes of any recognizable conditional mean function. Nor are the derivatives of the possibly interesting  $\text{Prob}(d_{i,1} = 1, d_{i,2} = 1 \mid \mathbf{w}_{i,1}, \mathbf{w}_{i,2})$ . Both of these are complicated functions of all the model parameters and both data vectors. [See Greene (2008, Section 23.8.3 and Christofides et al. (1997, 2000). Since this is a two equation model, it is unclear what quantity should be analyzed when interpreting the coefficients in relation to partial effects. One possibility is the joint probability,  $\text{Prob}(d_{i,1}=1, d_{i,2}=1) = \Phi_2[\mathbf{w}'_{i,1}\boldsymbol{\theta}_1, \mathbf{w}'_{i,2}\boldsymbol{\theta}_2, \rho]$ , that is analyzed by Christofides et al. Greene (1996, 2008) considers, instead, the conditional mean function  $E[d_{i,1} \mid d_{i,2} = 1, \mathbf{w}_{i,1}, \mathbf{w}_{i,2}] = \Phi_2[\mathbf{w}'_{i,1}\boldsymbol{\theta}_1, \mathbf{w}'_{i,2}\boldsymbol{\theta}_2, \rho] / \Phi[\mathbf{w}'_{i,2}\boldsymbol{\theta}_2]$ . In either case, the raw coefficients bear little resemblance to the partial effects.

For hypothesis testing about the coefficients, the standard results for Wald, LM and LR tests apply. The LM test is likely to be cumbersome because the derivatives of the log likelihood function are complicated. The other two are straightforward. A hypothesis of interest is the null hypothesis that the correlation is zero. For testing

$$H_0: \rho = 0,$$

all three likelihood based procedures are straightforward. The application below demonstrates. The Lagrange multiplier statistic derived by Kiefer (1982) is

$$\lambda_{LM} = \frac{\left\{ \sum_{i=1}^n \left[ q_{i,1} q_{i,2} \frac{\phi(\mathbf{w}'_{i,1} \boldsymbol{\theta}_1) \phi(\mathbf{w}'_{i,2} \boldsymbol{\theta}_2)}{\Phi(q_{i,1} \mathbf{w}'_{i,1} \boldsymbol{\theta}_1) \Phi(q_{i,2} \mathbf{w}'_{i,2} \boldsymbol{\theta}_2)} \right] \right\}^2}{\sum_{i=1}^n \left\{ \frac{[\phi(\mathbf{w}'_{i,1} \boldsymbol{\theta}_1) \phi(\mathbf{w}'_{i,2} \boldsymbol{\theta}_2)]^2}{\Phi(\mathbf{w}'_{i,1} \boldsymbol{\theta}_1) \Phi(-\mathbf{w}'_{i,1} \boldsymbol{\theta}_1) \Phi(\mathbf{w}'_{i,2} \boldsymbol{\theta}_2) \Phi(-\mathbf{w}'_{i,2} \boldsymbol{\theta}_2)} \right\}}$$

where the two coefficient vectors are the MLEs from the univariate probit models estimated separately.

#### 0.4.2 Recursive simultaneous equations

Section 0.3.5 considered a type of simultaneous equations model in which an endogenous regressor appears on the right hand side of a probit model. Two other simultaneous equations specifications have attracted interest. Amemiya (1985) demonstrates that a fully simultaneous bivariate probit model,

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}' \boldsymbol{\theta}_1 + \gamma_1 d_{i,2} + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}' \boldsymbol{\theta}_2 + \gamma_2 d_{i,1} + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, \end{aligned}$$

is internally inconsistent, and unidentified. However, a recursive model,

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}' \boldsymbol{\theta}_1 + \varepsilon_{i,1}, \quad d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}' \boldsymbol{\theta}_2 + \gamma_2 d_{i,1} + \varepsilon_{i,2}, \quad d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, \\ (\varepsilon_{i,1} \varepsilon_{i,2}) &\sim N_2[(0,0), (1,1), \rho] \end{aligned}$$

is a straightforward extension of the model. (For estimation of this model we have the counterintuitive result that it can be fit as an ordinary bivariate probit model with the additional right hand side variable in the second equation, ignoring the simultaneity. The recent literature provides a variety of applications of this model including Greene (1998), Fabbri, Monfardini and Radice (2004), Kassouf and Hoffman (2006), White and Wolaver (2003), Gandelman (2005), and Greene et al. (2006).

Interpretation of the components of this model is particularly complicated. Typically, interest will center on the second equation. For a few of the examples cited, in Greene (1998), the second equation concerned presence of a gender economics course in a college curriculum while the first equation specified presence of a women's studies program on the campus. In Kassouf and Hoffman (2003), the authors were interested in the occurrence of work related injuries while the first, conditioning equation specified the use or nonuse of protective equipment. Fabbri et al. (2004) analyzed the choice of Cesarean delivery conditioned on hospital type (public or private). In Greene et al. (2006), the main equation concerned use of a check cashing facility

while the conditioning event in the first equation was whether or not the individual participated in the banking system. In all of these cases, the margin of interest is the impact of the variables in the model on the probability that  $d_{i,2}$  equals one. Because  $d_{i,1}$  appears in the equation, there is (potentially) a direct effect (in  $\mathbf{w}_{i,2}$ ) and an indirect effect transmitted to  $d_{i,2}$  through the impact of the variable in question on the probability that  $d_{i,1}$  equals one. Details on these computations appear in Greene (2008) and Kassouf and Hoffmann (2006).

### 0.4.3 Sample selection in a bivariate probit model

Another bivariate probit model that is related to the recursive model of the preceding section is the *bivariate probit with sample selection*. The structural equations are

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, 0 \text{ otherwise,} \\ d_{i,2}^* &= \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, d_{i,2} = 1 \text{ if } d_{i,2}^* > 0, 0 \text{ otherwise, and if } d_{i,1} = 1, \\ d_{i,2}, \mathbf{w}_{i,2} &\text{ are unobserved when } d_{i,1} = 0, \\ (\varepsilon_{i,1}, \varepsilon_{i,2}) &\sim N_2[(0,0), (1,1), \rho]. \end{aligned}$$

The first equation is a ‘selection equation.’ Presence in the sample for observation of the second equation is determined by the first. Like the recursive model, this framework has been used in a variety of applications. The first was a study of the choice of deductibles in insurance coverage by Wynand and van Praag (1981). Boyes, Hoffman and Low (1989) and Greene (1992) studied loan default in which the application is the selection rule. More recently, McQuestion (2000) has used the model to analyze health status (selection) and health behavior, and Lee et al. (2003) have studied consumer adoption of computer banking technology.

Estimation of this sample selection model is done by maximum likelihood in one step.<sup>21</sup> The log likelihood is

$$\ln L = \sum_{d_{i,a}=0} \ln \Phi(-\mathbf{w}'_{i,1}\boldsymbol{\theta}_1) + \sum_{i=1, d_{i,1}=1}^n \ln \Phi_2[\mathbf{w}'_{i,1}\boldsymbol{\theta}_1, q_{i,2}\mathbf{w}'_{i,2}\boldsymbol{\theta}_2, q_{i,2}\rho]$$

As before, estimation and inference in this model follows the standard procedures.

### 0.4.4 Multivariate binary choice and the panel probit model

In principle, the bivariate probit model can be extended to an arbitrary number of equations, as

$$\begin{aligned} d_{i,1}^* &= \mathbf{w}_{i,1}'\boldsymbol{\theta}_1 + \varepsilon_{i,1}, d_{i,1} = 1 \text{ if } d_{i,1}^* > 0, \\ d_{i,2}^* &= \mathbf{w}_{i,2}'\boldsymbol{\theta}_2 + \varepsilon_{i,2}, d_{i,2} = 1 \text{ if } d_{i,2}^* > 0 \\ &\dots \\ d_{i,M} &= \mathbf{w}_{i,M}'\boldsymbol{\theta}_M + \varepsilon_{i,M}, d_{i,M} = 1 \text{ if } d_{i,M}^* > 0, \end{aligned}$$

$$\begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \dots \\ \varepsilon_{i,M} \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1M} \\ \rho_{12} & 1 & \dots & \rho_{2M} \\ \dots & \dots & \dots & \dots \\ \rho_{1M} & \rho_{2M} & \dots & 1 \end{pmatrix} \right] = N_M[\mathbf{0}, \mathbf{R}].$$

<sup>21</sup> Wynand and van Praag used a two step procedure similar to Heckman/s (1979) procedure for the linear model. Applications since then have used the MLE.

The obstacle to use of this model is the computational burden. The log likelihood is computed as follows: Let

$$\begin{aligned}\mathbf{Q}_i &= \text{diag}(q_{i,1}, q_{i,2}, \dots, q_{i,M}) \\ \mathbf{b}_i &= (\mathbf{w}_{i,1}'\boldsymbol{\theta}_1, \mathbf{w}_{i,2}'\boldsymbol{\theta}_2, \dots, \mathbf{w}_{i,M}'\boldsymbol{\theta}_M)' \\ \mathbf{c}_i &= \mathbf{Q}_i\mathbf{b}_i \\ \mathbf{D}_i &= \mathbf{Q}_i\mathbf{R}\mathbf{Q}_i\end{aligned}$$

then,

$$\ln L = \sum_{i=1}^n \ln \Phi_M[\mathbf{c}_i, \mathbf{D}_i]$$

Evaluation of the  $M$  variate normal cdf cannot be done analytically or with quadrature. It is done with simulation, using the GHK simulator. [See Geweke, et al. (1994).]

This form of the model also generalizes the random effects probit model examined earlier. We can relax the assumption of equal cross period correlations by writing

$$\begin{aligned}d_{it}^* &= \mathbf{w}_{it}'\boldsymbol{\theta} + \varepsilon_{it}, \quad d_{it} = 1 \text{ if } d_{it}^* > 0, 0 \text{ otherwise,} \\ (\varepsilon_{i1}, \dots, \varepsilon_{iT}) &\sim N[\mathbf{0}, \mathbf{R}].\end{aligned}$$

This is precisely the model immediately above with the constraint that the coefficients in the equations are all the same. In this form, it is conventionally labeled the *panel probit model*.<sup>22</sup> Bertschek and Lechner (1998) devised a GMM estimator to circumvent the computational burden of this model. Greene (2004a) examined the same model, and considered alternative computational procedures as well some variations on the model specification.

#### 0.4.5 Application

Riphahn et al. studied the joint determination of two counts, doctor visits and hospital visits. One would expect these to be highly correlated, so a bivariate probit model should apply to  $Doctor = 1(DoctorVis > 0)$  and  $Hospital = 1(HospitalVis > 0)$ . The simple product moment correlation coefficient is inappropriate for binary variables. The *tetrachoric correlation* is used instead; this turns out to be the estimate of  $\rho$  in a bivariate probit model in which both equations contain only a constant term. The first estimated model in Table 4 reports a value of 0.311 with a standard error of only 0.0136, so the results are consistent with the conjecture. The second estimates assume  $\rho = 0$ ; those for the ‘Doctor’ equation are reproduced from Table 3. As noted, there is evidence that  $\rho$  is positive. Kiefer’s Lagrange multiplier statistic equals 399.20. The limiting distribution is chi squared with one degree of freedom – the critical value is 3.84, so the hypothesis that the outcomes are conditionally uncorrelated is rejected. The Wald and likelihood ratio statistics based on the unrestricted model are  $21.496^2 = 462.08$  and  $2[17670.94 + 8084.465 - 25534.46] = 441.998$ , respectively, so the hypothesis is rejected based on all three tests. The third model in Table 4 is the unrestricted bivariate probit model. The fourth model shown in Table 4 is the recursive bivariate probit model with *Doctor* added to the right hand side of the *Hospital* equation. The results do not support this specification; the log likelihood is almost unchanged. It is noteworthy that in this expanded specification, the estimate of  $\rho$  is no longer significant, as might have been expected.

<sup>22</sup> Since the coefficient vectors are assumed to be the same in every period, it is only necessary to normalize one of the diagonal elements in  $\mathbf{R}$  to 1.0. See Greene (2004a) for discussion.

**Table 4 Estimated Bivariate Probit Models (Standard errors in parentheses)**

	(1) Tetrachoric Corr.		(2) Uncorrelated		(3) Bivariate Probit		(4) Recursive Probit		
	Doctor	Hospital	Doctor	Hospital	Doctor	Hospital	Doctor	Hospital	
Constant	.329 (.0077)	-1.355 (.0107)	.155 (.0565)	-1.246 (.0809)	.155 (.0565)	-1.249 (.0773)	.155 (.0565)	-1.256 (.481)	
Age	.000 (.000)	.000 (.000)	.0128 (.0008)	.00488 (.0011)	.0128 (.0008)	.00489 (.0011)	.0128 (.0008)	.00486 (.0025)	
Hhninc	.000 (.000)	.000 (.000)	-.116 (.0463)	.0421 (.0633)	-.118 (.0462)	.0492 (.0595)	-.118 (.0463)	.0496 (.0652)	
Hhkids	.000 (.000)	.000 (.000)	-.141 (.0182)	-.0147 (.0256)	-.141 (.0181)	-.0129 (.0257)	-.141 (.0181)	-.0125 (.0386)	
Educ	.000 (.000)	.000 (.000)	-.0281 (.0035)	-.0260 (.0052)	-.0280 (.0035)	-.0260 (.0051)	-.0280 (.0035)	-.0260 (.0066)	
Married	.000 (.000)	.000 (.000)	.0522 (.0205)	-.0547 (.0279)	.0519 (.0205)	-.0546 (.0277)	.0519 (.0205)	-.0548 (.0313)	
Doctor									.00912 (.663)
$\rho$	.311 (.0136)		0.000		.303 (.0138)		.298 (.366)		
Ln L	-25898.27		-17670.94	-8084.47	-25534.46		-25534.46		

**0.5 Ordered choice**

In the preceding sections, the consumer is assumed to maximize utility over a pair of alternatives. Models of ordered choice describe settings in which individuals reveal the strength of their utility with respect to a single outcome. For example, in a survey of voter preferences over a single issue (a new public facility or project, a political candidate, etc.), random utility is, as before,

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

The individual reveals a censored version of  $U_i^*$  through a discrete response, for example,

- $y_i = 0$ : strongly dislike,
- 1: mildly dislike,
- 2: indifferent,
- 3: mildly prefer,
- 4: strongly prefer.

The translation between the underlying  $U_i^*$  and the observed  $y_i$  produces the ordered choice model,

$$y_i = \begin{matrix} 0 & \text{if } U_i^* \leq \mu_0 \\ 1 & \text{if } 0 < U_i^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < U_i^* \leq \mu_2 \\ \dots & \\ J & \text{if } \mu_{J-1} < U_i^* \leq \mu_J. \end{matrix}$$

where  $\mu_0, \dots, \mu_J$  are threshold parameters that are to be estimated with the other model parameters subject to  $\mu_j > \mu_{j-1}$  for all  $j$ . Assuming  $\boldsymbol{\beta}$  contains a constant term, the distribution is located by the normalization  $\mu_0 = 0$ . (Note that in the form of our initial specification, this model would not contain any choice specific  $\mathbf{x}_i$  variables, as there is only one ‘alternative.’ We retain the current notation for simplicity and consistency with other treatments.) At the upper tail,  $\mu_J = +\infty$ . Probabilities for the observed outcomes are derived from the laws of probability,

$$\begin{aligned}\text{Prob}(y_i = j | \mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(d_{i,j} = 1 | \mathbf{x}_i, \mathbf{z}_i) \\ &= \text{Prob}(\mu_{j-1} < U_i^* \leq \mu_j) \text{ where } \mu_{-1} = -\infty.\end{aligned}$$

As before, the observed data do not reveal information about the scaling of  $\varepsilon_i$ , so the variance is normalized to one. Two standard cases appear in the literature; if  $\varepsilon_i$  has a normal distribution, then the *ordered probit model* emerges while if it has the standardized logistic distribution, the *ordered logit model* is produced. (Other distributions have been suggested – the model is internally consistent with any continuous distribution over the real line – however, these two overwhelmingly dominate the received applications.)

By the laws of probability,

$$\begin{aligned}\text{Prob}(y_i = j | \mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(U_i^* \leq \mu_j) - \text{Prob}(U_i^* \leq \mu_{j-1}) \\ &= F(\mu_j - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma}) - F(\mu_{j-1} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma})\end{aligned}$$

where  $F(c)$  is the assumed cdf, either normal or logistic. These are the terms that enter the log likelihood for a sample of  $n$  observations. The standard conditions for maximum likelihood estimation apply here. The results in Table 1 suggest that the force of the incidental parameters problem in the fixed effects case is similar to that for the binomial probit model.

As usual in discrete choice models, partial effects in this model differ substantively from the coefficients. Note, first, there is no obvious regression at work. Since  $y_i$  is merely a labeling with no implicit scale, there is no conditional mean function to analyze. In order to analyze the impact of changes in a variable, say income, one can decompose the set of probabilities. For a continuous variable in  $\mathbf{x}_i$ , for example,

$$\delta_{i,k}(j) = \partial \text{Prob}(y_i = j | \mathbf{x}_i, \mathbf{z}_i) / \partial x_{i,k} = -\beta_k [f(\mu_j - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma}) - f(\mu_{j-1} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma})], j = 0, \dots, J,$$

where  $f(c)$  is the density,  $dF(c)/dc$ . The sign of the partial effect is ambiguous, since the difference of the two densities can have either sign. Moreover, since  $\sum_{j=0}^J \text{Prob}(y_i = j | \mathbf{x}_i, \mathbf{z}_i) = 1$ , it follows that  $\sum_{j=1}^J \delta_{i,k}(j) = 0$ . Since the cdf is monotonic, there is one sign change in the set of partial effects. The example below demonstrates. For purposes of using and interpreting the model, it seems that the coefficients are of relatively little utility – neither the sign nor the magnitude directly indicates the effect of changes in a variable on the observed outcome.

Terza (1985) and Pudney and Shields (2000) suggested an extension of the ordered choice model that would accommodate heterogeneity in the threshold parameters. The extended model is

$$\text{Prob}(y_i = j | \mathbf{x}_i, \mathbf{z}_i) = F(\mu_{i,j} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma}) - F(\mu_{i,j-1} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma})$$

where

$$\mu_{i,j} = \mathbf{v}_i' \boldsymbol{\pi}_j \text{ where } \boldsymbol{\pi}_0 = \mathbf{0}$$

for a set of variables  $\mathbf{v}_i$ . The model as shown has two complications. First, it is straightforward to constrain the fixed threshold parameters to preserve the ordering needed to ensure that all probabilities are positive.<sup>23</sup> When there are variables  $\mathbf{v}_i$  in the construction, it is no longer possible to produce this result parametrically. The authors (apparently) did not find it necessary to confront this constraint. As second feature of the model (which was examined at length by the

<sup>23</sup> For example, the parameters can be written in terms of a set of latent parameters so that  $\mu_1 = \tau_1^2$ ,  $\mu_2 = \tau_1^2 + \tau_2^2$ , and so on. Typically, the explicit reparameterization is unnecessary.

authors) is the unidentifiability of elements of  $\boldsymbol{\pi}_j$  when  $\mathbf{v}_i$  and  $(\mathbf{x}_i, \mathbf{z}_i)$  contain the same variables. This is a result of the linear functional form assumed for  $\mu_{i,j}$ . Greene (2007), Harris and Zhao (2007) and Greene et al. (2007) suggested alternative parameterizations that circumvents these problems, a restricted version,

$$\mu_{i,j} = \exp(\mu_j + \mathbf{v}_i' \boldsymbol{\pi})$$

and a counterpart to Pudney and Shields's formulation,

$$\mu_{i,j} = \exp(\mathbf{v}_i' \boldsymbol{\pi}_j).^{24}$$

### 0.5.1 Specification analysis

As in the binary choice case, the analysis of micro- level data is likely to encounter individual heterogeneity not only in the means of utilities  $(\mathbf{x}_i, \mathbf{z}_i)$  but also in the scaling of  $U_i^*$ , that is, in the variance of  $\varepsilon_i$ . Building heteroscedasticity into the model as in the binary choice model shown earlier is straightforward. If

$$E[\varepsilon_i^2 | \mathbf{v}_i] = [\exp(\mathbf{v}_i' \boldsymbol{\tau})]^2$$

then the log likelihood would become

$$\ln L = \sum_{i=1}^n \ln \left[ F \left( \frac{\mu_{y_i} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma}}{\exp(\mathbf{v}_i' \boldsymbol{\tau})} \right) - F \left( \frac{\mu_{y_i-1} - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{z}_i' \boldsymbol{\gamma}}{\exp(\mathbf{v}_i' \boldsymbol{\tau})} \right) \right].$$

As before, this complicates (even further) the interpretation of the model components and the partial effects.

There is no direct test for the distribution, since the alternatives are not nested. The Vuong test is a possibility, however the power of this test and its characteristics remain to be examined both analytically and empirically.

### 0.5.2 Bivariate ordered probit models

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 0.4. A direct analog to the base case two equation model was used by Butler et al. (1998) who analyzed the relationship between the level of calculus attained and grades in intermediate Economics courses for a sample of Vanderbilt University students. The two step estimation approach involved the following strategy: (We are stylizing the precise formulation a bit in order to compress the description.) Step 1 involved a direct application of the

---

<sup>24</sup> One could argue that this reformulation achieves identification purely 'through functional form' rather than through the theoretical underpinnings of the model. Of course, this assertion elevates the linear specification to a default position of prominence which seems unwarranted. Moreover, arguably, the underlying theory (as in fact suggested in passing by Pudney and Shields (2000)) is that there are different effects of the regressors on the thresholds and on the underlying utility.

ordered probit model of Section 0.5 to the level of calculus achievement, which is coded 0,1,...,6;

$$\begin{aligned}
 m_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i | \mathbf{x}_i \sim N[0,1], \\
 m_i &= 0 \text{ if } -\infty < m_i^* \leq 0, \\
 &1 \text{ if } 0 < m_i^* \leq \mu_1, \\
 &\dots \\
 &6 \text{ if } \mu_5 < m_i^* < +\infty.
 \end{aligned}$$

The authors argued that although the various calculus courses can be ordered discretely by the material covered, the differences between the levels cannot be measured directly. Thus, this is an application of the ordered probit model. The independent variables in this first step model included SAT scores, foreign language proficiency, indicators of intended major and several other variables related to areas of study.

The second step of the estimator involves regression analysis of the grade in the intermediate microeconomics or macroeconomics course. Grades in these courses were translated to a granular continuous scale (A = 4.0, A- = 3.7, etc.). A linear regression is specified,

$$Grade_i = \mathbf{z}_i' \boldsymbol{\delta} + u_i, \text{ where } u_i | \mathbf{z}_i \sim N[0, \sigma_u^2].$$

Independent variables in this regression include, among others, (1) dummy variables for which outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman GPA, etc. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that  $(\varepsilon_i, u_i | \mathbf{x}_i, \mathbf{z}_i) \sim N_2[(0,0), (1, \sigma_u^2), \rho \sigma_u]$ . A nonzero  $\rho$  captures this “selection” effect. With this in place, the dummy variables in (1) above have now become endogenous. The solution is a *selection correction*,

$$\begin{aligned}
 Grade_i | m_i &= \mathbf{z}_i' \boldsymbol{\delta} + E[u_i | m_i] + v_i \\
 &= \mathbf{z}_i' \boldsymbol{\delta} + (\rho \sigma_u) [\lambda(\mathbf{x}_i' \boldsymbol{\beta}, \mu_1, \dots, \mu_5)] + v_i.
 \end{aligned}$$

They thus adopt a control function approach to accommodate the endogeneity of the math attainment dummy variables. The term  $\lambda(\mathbf{x}_i' \boldsymbol{\beta}, \mu_1, \dots, \mu_5)$  is a *generalized residual* that is constructed using the estimates from the first stage ordered probit model. [A precise statement of the form of this variable is given in Tobias and Li (2006).] Linear regression of the course grade on  $\mathbf{z}_i$  and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (1985) correction.

Li and Tobias (2006) in a replication of and comment on Butler et al. (1998), after roughly replicating the classical estimation results with a Bayesian estimator, observe that the *Grade* equation above could also be treated as an ordered probit model. The resulting *bivariate ordered probit* model would be

$$\begin{aligned}
 m_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, & \text{and} & & g_i^* &= \mathbf{z}_i' \boldsymbol{\delta} + u_i, \\
 m_i &= 0 \text{ if } -\infty < m_i^* \leq 0, & & & g_i &= 0 \text{ if } -\infty < g_i^* \leq 0, \\
 &1 \text{ if } 0 < m_i^* \leq \mu_1, & & & &1 \text{ if } 0 < g_i^* \leq \alpha_1, \\
 &\dots & & & &\dots \\
 &6 \text{ if } \mu_5 < m_i^* < +\infty. & & & &11 \text{ if } \mu_9 < g_i^* < +\infty
 \end{aligned}$$

where  $(\varepsilon_i, u_i | \mathbf{x}_i, \mathbf{z}_i) \sim N_2[(0,0), (1, \sigma_u^2), \rho \sigma_u]$ .

Tobias and Li extended their analysis to this case simply by “transforming” the dependent variable in Butler et al.’s second equation. Computing the log likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model. [See Greene (2007).] However, the classical study of these data using the bivariate ordered approach remains to be done, so a side by side comparison to Tobias and Li’s Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables in (1) remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appear.

The bivariate ordered probit model has been applied in a number of settings in the recent empirical literature, including husband and wife’s education levels [Magee et al. (2000)], family size [(Calhoun (1991))] and many others. In two early contributions to the field of pet econometrics, Butler and Chatterjee analyze ownership of cats and dogs (1995) and dogs and televisions (1997).

### 0.5.3 Panel data applications

#### *Fixed effects*

D’Addio et al. (2003), using methodology developed by Frijters et al. (2004) and Ferrer-i-Carbonel et al. (2004) analyzed survey data on job satisfaction using the Danish component of the European Community Household Panel. Their estimator for an ordered *logit* model is built around the logic of Chamberlain’s estimator for the binary logit model. (Section 23.5.2). Since the approach is robust to individual specific threshold parameters and allows time invariant variables, so it differs sharply from the fixed effects models we have considered thus far as well as from the ordered probit model of Section 0.5. Unlike Chamberlain’s estimator for the binary logit model, however, their conditional estimator is not a function of minimal sufficient statistics. As such, the incidental parameters problem remains an issue.

Das and van Soest (1999) proposed a somewhat simpler approach. [See, as well, Long’s (1997) discussion of the “parallel regressions assumption,” which employs this device in a cross section framework.] Consider the base case ordered logit model with fixed effects,

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it}|\mathbf{X}_i \sim N[0,1]$$

$$y_{it} = j \text{ if } \mu_{j-1} < y_{it}^* < \mu_j, j = 0, 1, \dots, J \text{ and } \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that

$$\text{Prob}(y_{it} = j|\mathbf{X}_i) = \Lambda(\mu_j - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) - \Lambda(\mu_{j-1} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta})$$

where  $\Lambda(c)$  is the cdf of the logistic distribution. Now, define a binary variable

$$w_{it,j} = 1 \text{ if } y_{it} > j, j = 0, \dots, J-1.$$

It follows that

$$\begin{aligned} \text{Prob}[w_{it,j} = 1|\mathbf{X}_i] &= \Lambda(\alpha_i - \mu_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= \Lambda(\theta_i + \mathbf{x}_{it}'\boldsymbol{\beta}). \end{aligned}$$

The “ $j$ ” specific constant, which is the same for all individuals, is absorbed in  $\theta_i$ . Thus, a fixed effects binary logit model applies to each of the  $J-1$  binary random variables,  $w_{it,j}$ . The method in Section 0.3.6 can now be applied to each of the  $J-1$  random samples. 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  $\beta$ . The minimum distance estimator at the second step is chosen to minimize

$$q = \sum_{j=0}^{J-1} \sum_{m=0}^{J-1} (\hat{\beta}_j - \beta)' [V_{jm}^{-1}] (\hat{\beta}_m - \beta)$$

where  $[V_{jm}^{-1}]$  is the  $j,m$  block of the inverse of the  $(J-1)K \times (J-1)K$  partitioned matrix  $V$  that contains  $\text{Asy.Cov}[\hat{\beta}_j, \hat{\beta}_m]$ . The appropriate form of this matrix for a set of cross section estimators is given in Brant (1990). Das and van Soest (2000) used the counterpart for Chamberlain's fixed effects estimator, but do not provide the specifics for computing the off diagonal blocks in  $V$ .

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 (2008, Section 16.9.6.c). The likelihood function is concave [see Pratt (1981)], so despite its superficial complexity, the estimation is straightforward. (In the application below, with over 27,000 observations and 7,293 individual effects, estimation of the full model required roughly five seconds of computation.) No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) results 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 Table 1, suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. As in the binary choice case, the complication of the fixed effects model is the small sample 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.

#### Random effects

The random effects ordered probit model has been much more widely used than the fixed effects model. Applications include Groot and van den Brink (2003) who studied training levels of employees, with firm effects and gains to marriage, Winkelmann (2004) who examined subjective measures of well being with individual and family effects, Contoyannis et al. who analyzed self reported measures of health status and numerous others. In the simplest case, the quadrature method of Butler and Moffitt (1982) can be used to maximize the log likelihood.

#### 0.5.4 Applications

The German Health Care data that we have used earlier includes a self reported measure of health satisfaction, *HSAT*, that takes values 0,1,...,10. This is a typical application of a scale variable that reflects an underlying continuous variable, "health." The frequencies and sample proportions for the reported values are as follows:

<i>NEWHSAT</i>	0	1	2	3	4	5	6	7	8	9	10
<i>Frequency</i>	447	255	642	1173	1390	4233	2530	4231	6172	3061	3192
<i>Proportion</i>	1.6%	0.9%	2.3%	4.2%	5.0%	15.4%	9.2%	15.4%	22.5%	11.2%	11.6%

We have fit pooled, and panel data versions of the ordered probit model to these data. The model used is

$$U_{it}^* = \beta_1 + \beta_2 \text{Age}_{it} + \beta_3 \text{Income}_{it} + \beta_4 \text{Education}_{it} + \beta_5 \text{Married}_{it} + \beta_6 \text{Working}_{it} + \varepsilon_{it} + c_i$$

where  $c_i$  will be the common fixed or random effect. Table 5 lists five estimated models. (Standard errors for the estimated threshold parameters are omitted.) The first is the pooled ordered probit model. The second and third are fixed effects. Column 2 shows the unconditional fixed effects estimates using the results in Greene (2008). Column 3 shows the Das and van Soest estimator. For the minimum distance estimator, we used an inefficient weighting matrix, the block diagonal matrix in which the  $j$ th block is the inverse of the  $j$ th asymptotic covariance matrix for the individual logit estimators. With this weighting matrix, the estimator is

$$\hat{\beta}_{MDE} = \left[ \sum_{j=0}^9 \mathbf{V}_j^{-1} \right]^{-1} \sum_{j=0}^9 \mathbf{V}_j^{-1} \hat{\beta}_j$$

and the estimator of the asymptotic covariance matrix is approximately equal to the bracketed inverse matrix. The fourth set of results is the random effects estimator computed using the maximum simulated likelihood method. This model can be estimated using Butler and Moffitt's quadrature method, however we found that even with a large number of nodes, the quadrature estimator converged to a point where the log likelihood was far lower than the MSL estimator, and at parameter values that were implausibly different from the other estimates. Using different starting values and different numbers of quadrature points did not change this outcome. The MSL estimator for a random constant term is considerably slower, but produces more reasonable results. The fifth set of results is the Mundlak form of the random effects model, that includes the group means in the models as controls to accommodate possible correlation between the latent heterogeneity and the included variables. As noted earlier the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 6 shows the partial effects for the pooled model to illustrate the computations.

Winkelmann (2004) used the random effects approach to analyze the subjective well being (SWB) question (also coded 0 to 10) in the German Socioeconomic Panel (GSOEP) data set. The ordered probit model in this study is based on the latent regression

$$y_{imt}^* = \mathbf{x}_{imt}'\boldsymbol{\beta} + \varepsilon_{imt} + u_{im} + v_i.$$

The independent variables include age, gender, employment status, income, family size and an indicator for good health. An unusual feature of the model is the nested random effects [see Greene (2008, Section 9.7.1)] which include a family effect,  $v_i$ , as well as the individual family member ( $i$  in family  $m$ ) effect,  $u_{im}$ . The GLS/MLE approach that would be used for the linear regression model is unavailable in this nonlinear setting. Winkelmann, instead employed Hermite quadrature procedure to maximize the log likelihood function.

Contoyannis, Jones and Rice (2004) analyzed a self assessed health scale that ranged from 1 (very poor) to 5 (excellent) in the British Household Panel Survey. Their model accommodated a variety of complications in survey data. The latent regression underlying their ordered probit model is

$$h_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{H}_{i,t-1}'\boldsymbol{\gamma} + \alpha_i + \varepsilon_{it},$$

where  $\mathbf{x}_{it}$  includes marital status, race, education, household size, age, income, and number of children in the household. The lagged value,  $\mathbf{H}_{i,t-1}$  is a set of binary variables for the observed health status in the previous period. ) In this case, the lagged values capture state dependence – the assumption that the health outcome is redrawn randomly in each period is inconsistent with evident runs in the data. The initial formulation of the regression is a fixed effects model. To control for the possible correlation between the effects,  $\alpha_i$ , and the regressors, and the initial

conditions problem that helps to explain the state dependence, they use a hybrid of Mundlak's (1978) correction and a suggestion by Wooldridge (2002) for modeling the initial conditions,

$$\alpha_i = \alpha_0 + \bar{x}'\alpha_1 + \mathbf{H}_{i,1}'\delta + u_i,$$

where  $u_i$  is exogenous. Inserting the second equation into the first produces a random effects model that can be fit using Butler and Moffitt's quadrature method.

**Table 5 Estimated Ordered Probit Models for Health Satisfaction**

<i>Variable</i>	(1)	(2)	(3)	(4)	(5)	
	Pooled	Fixed Effects Unconditional	Fixed Effects Conditional	Random Effects	Random Effects Mundlak Controls Variables	Means
<i>Constant</i>	2.4739 (0.04669)			3.8577 (0.05072)	3.2603 (0.05323)	
<i>Age</i>	-0.01913 (0.00064)	-0.07162 (0.002743)	-0.1011 (0.002878)	-0.03319 (0.00065)	-0.06282 (0.00234)	0.03940 (0.002442)
<i>Income</i>	0.1811 (0.03774)	0.2992 (0.07058)	0.4353 (0.07462)	0.09436 (0.03632)	0.2618 (0.06156)	0.1461 (0.07695)
<i>Kids</i>	0.06081 (0.01459)	-0.06385 (0.02837)	-0.1170 (0.03041)	0.01410 (0.01421)	-0.05458 (0.02566)	0.1854 (0.03129)
<i>Education</i>	0.03421 (0.002828)	0.02590 (0.02677)	0.06013 (0.02819)	0.04728 (0.002863)	0.02296 (0.02793)	0.02257 (0.02807)
<i>Married</i>	0.02574 (0.01623)	0.05157 (0.04030)	0.08505 (0.04181)	0.07327 (0.01575)	0.04605 (0.03506)	-0.04829 (0.03963)
<i>Working</i>	0.1292 (0.01403)	-0.02659 (0.02758)	-0.007969 (0.02830)	0.07108 (0.01338)	-0.02383 (0.02311)	0.2702 (0.02856)
$\mu_1$	0.1949	0.3249		0.2726	0.2752	
$\mu_2$	0.5029	0.8449		0.7060	0.7119	
$\mu_3$	0.8411	1.3940		1.1778	1.1867	
$\mu_4$	1.111	1.8230		1.5512	1.5623	
$\mu_5$	1.6700	2.6992		2.3244	2.3379	
$\mu_6$	1.9350	3.1272		2.6957	2.7097	
$\mu_7$	2.3468	3.7923		3.2757	3.2911	
$\mu_8$	3.0023	4.8436		4.1967	4.2168	
$\mu_9$	3.4615	5.5727		4.8308	4.8569	
$\sigma_u$	0.0000	0.0000		1.0078	0.9936	
$\ln L$	-56813.52	-41875.63		-53215.54	-53070.43	

**Table 6 Estimated Marginal Effects: Pooled Model**

HSAT	Age	Income	Kids	Education	Married	Working
0	0.0006	-0.0061	-0.0020	-0.0012	-0.0009	-0.0046
1	0.0003	-0.0031	-0.0010	-0.0006	-0.0004	-0.0023
2	0.0008	-0.0072	-0.0024	-0.0014	-0.0010	-0.0053
3	0.0012	-0.0113	-0.0038	-0.0021	-0.0016	-0.0083
4	0.0012	-0.0111	-0.0037	-0.0021	-0.0016	-0.0080
5	0.0024	-0.0231	-0.0078	-0.0044	-0.0033	-0.0163
6	0.0008	-0.0073	-0.0025	-0.0014	-0.0010	-0.0050
7	0.0003	-0.0024	-0.0009	-0.0005	-0.0003	-0.0012
8	-0.0019	0.0184	0.0061	0.0035	0.0026	0.0136
9	-0.0021	0.0198	0.0066	0.0037	0.0028	0.0141
10	-0.0035	0.0336	0.0114	0.0063	0.0047	0.0233

## 0.6 Models for counts

A model that is often used for interarrival times at such facilities as telephone switches, ATM machines, or the service windows of banks or gasoline stations is the *exponential model*,

$$f(t) = \theta \exp(-\theta t), t \geq 0, \theta > 0,$$

where the continuous variable,  $t$ , is the time between arrivals. The expected interarrival time in this distribution is  $E[t] = 1/\theta$ . Consider the number of arrivals,  $y$ , that occur *per unit of time*. It can be shown that this discrete random variable has the *Poisson probability distribution*

$$f(y) = \exp(-\lambda)\lambda^y / y!, \lambda = 1/\theta > 0, y = 0, 1, \dots$$

The expected value of this discrete random variable is  $E[y] = 1/\theta$ . The *Poisson regression model* arises from the specification

$$E[y_i | \mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}).$$

The loglinear form is used to ensure that the mean is positive. Estimation of the Poisson model by maximum likelihood is straightforward owing to the simplicity of the log likelihood and its derivatives,

$$\begin{aligned} \ln L &= \sum_{i=1}^n -\lambda_i + y_i(\mathbf{x}_i' \boldsymbol{\beta}) - \ln \Gamma(y_i + 1) \\ \partial \ln L / \partial \boldsymbol{\beta} &= \sum_{i=1}^n (y_i - \lambda_i) \mathbf{x}_i \\ \partial^2 \ln L / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' &= \sum_{i=1}^n -\lambda_i \mathbf{x}_i \mathbf{x}_i'. \end{aligned}$$

Inference about parameters is based on either the actual (and expected) Hessian,

$$\mathbf{V} = \left[ \sum_{i=1}^n \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = \left[ \mathbf{X}' \hat{\boldsymbol{\Lambda}} \mathbf{X} \right]^{-1}$$

or the BHHH estimator which is

$$\mathbf{V}_{\text{BHHH}} = \left[ \sum_{i=1}^n (y_i - \hat{\lambda}_i)^2 \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = \left[ \sum_{i=1}^n \varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i' \right]^{-1} = \left[ \mathbf{X}' \hat{\mathbf{E}}^2 \mathbf{X} \right]^{-1}.$$

Hypothesis tests about the parameters may be based on the likelihood ratio or Wald statistics, or the Lagrange multiplier statistic which is particularly convenient here,

$$\lambda_{\text{LM}} = \left[ \sum_{i=1}^n \hat{\varepsilon}_i^0 \mathbf{x}_i \right]' \mathbf{V}_{\text{BHHH}} \left[ \sum_{i=1}^n \hat{\varepsilon}_i^0 \mathbf{x}_i \right]$$

where the residuals are computed at the restricted estimates. For example, under the null hypothesis that all coefficients are zero save for the constant term,  $\hat{\lambda}_i^0 = \bar{y}$ ,  $\hat{\varepsilon}_i^0 = y_i - \bar{y}$  and

$$\lambda_{\text{LM}} = \left[ \sum_{i=1}^n (y_i - \bar{y}) \mathbf{x}_i \right]' \left[ \sum_{i=1}^n (y_i - \bar{y})^2 \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[ \sum_{i=1}^n (y_i - \bar{y}) \mathbf{x}_i \right].$$

The Poisson model is one in which the MLE is robust to certain misspecifications of the model, such as the failure to incorporate latent heterogeneity in the mean (i.e., one fits the Poisson model when the negative binomial is appropriate.) In this case, the robust (sandwich) covariance matrix,

$$\text{Robust Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1} [\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}] [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1}$$

is appropriate to accommodate this failure of the model. It has become common to employ this estimator with all specifications, including the negative binomial. One might question the virtue of this. Since the negative binomial model already accounts for the latent heterogeneity, it is unclear what *additional* failure of the assumptions of the model this estimator would be robust to.

Since the model is a true regression model, the predicted values,  $\hat{\lambda}_i$  are meaningful quantitative forecasts of the actual outcomes,  $y_i$ . The simple squared correlation between  $\hat{\lambda}_i$  and  $y_i$  will be indicative of the fit of the model, though it is not a measure of variation explained. Moreover, though it is likely in practice, the theory does not guarantee that this  $r^2$ -like measure will increase when variables are added to the model. Several alternative measures of fit for count data models are suggested in Greene (2008, Chapter 25). One which has an intuitive appeal is based on the model deviances,  $e_i = 2[y_i \ln(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)]$  (where  $0 \ln 0$  is understood to equal zero). An associated fit measure based on this measure is

$$R_d^2 = \frac{l(\hat{\lambda}_i) - l(\bar{y})}{l(y_i) - l(\bar{y})},$$

where  $l(\hat{y}_i)$  indicates the evaluation of the log likelihood function with  $y_i$  predicted by the indicated  $\hat{y}_i$ . (Thus,  $l(y_i)$  is the original log likelihood.) [See Cameron and Windmeijer (1993).]

### 0.6.1 Heterogeneity and the negative binomial model

The Poisson model is typically only the departure point for the analysis of count data. The simple model has (at least) two shortcomings that arise from heterogeneity that is not explicitly incorporated in the model.

One easily remedied minor issue concerns the units of measurement of the data. In the Poisson model (and negative binomial model below), the parameter  $\lambda_i$  is the expected number of events *per unit of time*. Thus, there is a presumption in the model formulation, e.g., the Poisson, that the same amount of time is observed for each  $i$ . In a spatial context, such as measurements of the incidence of a disease per group of  $N_i$  persons, or the number of bomb craters per square mile (London, 1940), the assumption would be that the same physical area or the same size of population applies to each observation. Where this differs by individual, it will introduce a type of heteroscedasticity in the model. The simple remedy is to modify the model to account for the *exposure*,  $T_i$ , of the observation as follows:

$$\text{Prob}(y_i = j | \mathbf{x}_i, T_i) = \frac{\exp(-T_i\phi_i)(T_i\phi_i)^j}{j!}, \quad \phi_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad j = 0, 1, \dots$$

The original model is returned if we write  $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \ln T_i)$ . Thus, when the exposure differs by observation, the appropriate accommodation is to include the log of exposure in the regression part

of the model with a coefficient of 1.0. (For less than obvious reasons, the term “*offset variable*” is commonly associated with the exposure variable  $T_i$ .) Note that if  $T_i$  is the same for all  $i$ ,  $\ln T$  will simply vanish into the constant term of the model (assuming one is included in  $\mathbf{x}_i$ .)

The less straightforward restriction of the Poisson model is that  $E[y_i|\mathbf{x}_i] = \text{Var}[y_i|\mathbf{x}_i]$ . This *equidispersion* assumption is a major shortcoming. Observed data rarely if ever display this feature. The very large amount of research activity on functional forms for count models is often focused on testing for equidispersion and building functional forms that relax this assumption.

The overdispersion found in observed data can be attributed to omitted heterogeneity in the Poisson model. A more complete regression specification would be

$$E[y_i|\mathbf{x}_i] = \lambda_i = h_i \exp(\mathbf{x}_i'\boldsymbol{\beta}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)$$

where the heterogeneity,  $h_i$  has mean one and nonzero variance. Two candidates for the distribution of  $\varepsilon_i$  have dominated the literature, the lognormal model discussed later and the log gamma model. The more common specification is the log gamma model, which derives from the gamma variable,

$$f(h_i) = [\theta^\theta/\Gamma(\theta)]\exp(-\theta h_i)h_i^{\theta-1}, h_i \geq 0.^{25}$$

This gamma distributed random variable has mean 1.0 and variance  $1/\theta$ . (A separate variance parameter is not identified – the scaling in the model is, once again, absorbed by the coefficient vector.) If we write the Poisson – log gamma model as

$$f(y_i | \mathbf{x}_i, h_i) = \exp(-h_i \lambda_i)(h_i \lambda_i)^{y_i} / \Gamma(y_i + 1)$$

then the unconditional distribution is

$$f(y_i|\mathbf{x}_i) = \int_0^\infty f(y_i, h_i | \mathbf{x}_i) dv_i = \int_0^\infty f(y_i | \mathbf{x}_i, h_i) f(h_i) dh_i$$

The integral can be obtained in closed form; the result is the *negative binomial model*,

$$\begin{aligned} \text{Prob}(Y = y_i|\mathbf{x}_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i} (1 - r_i)^\theta, \\ \lambda_i &= \exp(\mathbf{x}_i'\boldsymbol{\beta}), \\ r_i &= \lambda_i / (\theta + \lambda_i). \end{aligned}$$

The recent literature, mostly associating the result with Cameron and Trivedi (1986, 1998) defines this form of the negative binomial model as the *Negbin 2* (NB2) form of the probability. This is the default form of the model in the common econometrics packages that provide an estimator for this model. The *Negbin 1* (NB1) form of the model results if  $\theta$  in the preceding is replaced with  $\theta_i = \theta \lambda_i$ . Then,  $r_i$  reduces to  $r = 1/(1+\theta)$ , and the density becomes

$$\text{Prob}(Y = y_i|\mathbf{x}_i) = \frac{\Gamma(\theta \lambda_i + y_i)}{\Gamma(y_i + 1)\Gamma(\theta \lambda_i)} r^{y_i} (1 - r)^{\theta \lambda_i}$$

---

<sup>25</sup> No theory justifies the choice of the log gamma density. It is essentially the same as a conjugate prior in Bayesian analysis, chosen for its mathematical convenience.

This is not a simple reparameterization of the model. The results in the example below demonstrate that the log likelihood functions are not equal at the maxima, and the parameters are not simple transformations in one model vs. the other. We are not aware of a theory that justifies using one form or the other for the negative binomial model. Neither is a restricted version of the other, so we cannot carry out a likelihood ratio test of one versus the other. The more general *Negbin P* (NBP) family [Greene (2007b)] does nest both of them, so this may provide a more general, encompassing approach to finding the right specification. The *Negbin P* model is obtained by replacing  $\theta$  in the *Negbin 2* form with  $\theta\lambda_i^{2-P}$ . We have examined the cases of  $P = 1$  and  $P = 2$  above.

$$\text{Prob}(Y = y_i | \mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i^Q + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i^Q)} \left( \frac{\lambda}{\theta\lambda_i^Q + \lambda_i} \right)^{y_i} \left( \frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i} \right)^{\theta\lambda_i^Q}, Q = 2 - P.$$

The conditional mean function for the three cases considered is

$$E[y_i | \mathbf{x}_i] = \exp(\mathbf{x}_i' \boldsymbol{\beta}) \times \theta^{2-P} = \alpha^{P-2} \lambda_i, \text{ where } \alpha = 1/\theta.$$

The parameter  $P$  is picking up the scaling. A general result is that for all three variants of the model,

$$\text{Var}[y_i | \mathbf{x}_i] = \lambda_i (1 + \alpha \lambda_i^{P-1}).$$

Thus, the NB2 form has a variance function that is quadratic in the mean while the NB1 form's variance is a simple multiple of the mean. There have been many other functional forms proposed for count data models, including the generalized Poisson, gamma, and Polya-Aeppli forms described in Winkelmann (2003) and Greene (2007a, Chapter 24).

The heteroscedasticity in the count models is induced by the relationship between the variance and the mean. The single parameter  $\theta$  picks up an implicit overall scaling, so it does not contribute to this aspect of the model. As in the linear model, microeconomic data are likely to induce heterogeneity in both the mean and variance of the response variable. A specification that allows independent variation of both will be of some virtue. The result

$$\text{Var}[y_i | \mathbf{x}_i] = \lambda_i (1 + (1/\theta)\lambda_i^{P-1})$$

suggests that a natural platform for separately modeling heteroscedasticity will be the dispersion parameter,  $\theta$ , which we now parameterize as

$$\theta_i = \theta \exp(\mathbf{z}_i' \boldsymbol{\delta})$$

Operationally, this is a relatively minor extension of the model. But, it is likely to introduce quite a substantial increase in the flexibility of the specification. Indeed, a heterogeneous *Negbin P* model is likely to be sufficiently parameterized to accommodate the behavior of most data sets. (Of course, the specialized models discussed below, for example, the zero inflation models, may yet be more appropriate for a given situation.)

## 0.6.2 Extended models for counts: Two part, zero inflation, sample selection, bivariate

Sources of 'nonPoissonness' arise from a variety of sources in addition to the latent heterogeneity modeled in the previous section. A variety of *two part models* have been proposed to accommodate elements of the decision process.

### Hurdle model

The hurdle model [Mullahy (1986), Gurmú (1997)] consists of a participation equation and a conditional Poisson or negative binomial model. The structural equations are

$$\begin{aligned} \text{Prob}(y_i > 0 \mid \mathbf{z}_i) &= \text{a binary choice mechanism, such as probit or logit} \\ \text{Prob}(y_i = j \mid y_i > 0, \mathbf{x}_i) &= \text{truncated Poisson or negative binomial.} \end{aligned}$$

[See Shaw (1988).] For a logit participation equation and a Poisson count, the probabilities for the observed data that enter the log likelihood function would be

$$\begin{aligned} \text{Prob}(y_i = 0 \mid \mathbf{z}_i) &= \frac{1}{1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})} \\ \text{Prob}(y_i = j \mid \mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(y_i > 0 \mid \mathbf{z}_i) \times \text{Prob}(y_i = j \mid y_i > 0, \mathbf{x}_i) \\ &= \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})} \frac{\exp(-\lambda_i) \lambda_i^j}{j! [1 - \exp(-\lambda_i)]}. \end{aligned}$$

This model might apply for on site counts of use of certain facilities such as recreation sites. The expectation in the hurdle model is easily found using the rules of probability

$$\begin{aligned} E[y_i \mid x_i, z_i] &= \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})} E[y_i \mid y_i > 0, \mathbf{x}_i, \mathbf{z}_i] \\ &= \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})} \frac{\lambda_i}{[1 - \exp(-\lambda_i)]}. \end{aligned}$$

As usual, the intricacy of the function mandates some caution in interpreting the model coefficients. In particular,

$$\begin{aligned} \delta_i(\mathbf{x}_i) &= \frac{\partial E[y_i \mid \mathbf{x}_i, \mathbf{z}_i]}{\partial x_i} \\ &= \left\{ \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})} \frac{\lambda_i}{[1 - \exp(-\lambda_i)]} \left( 1 - \frac{\lambda_i \exp(-\lambda_i)}{[1 - \exp(-\lambda_i)]} \right) \right\} \boldsymbol{\beta} \end{aligned}$$

The complication of the partial effects is compounded if  $\mathbf{z}_i$  contains any of the variables that also appear in  $\mathbf{x}_i$ . The other part of the partial effect is

$$\delta_i(\mathbf{z}_i) = \left\{ \frac{\exp(\mathbf{z}'_i \boldsymbol{\alpha})}{[1 + \exp(\mathbf{z}'_i \boldsymbol{\alpha})]^2} \frac{\lambda_i}{[1 - \exp(-\lambda_i)]} \right\} \boldsymbol{\alpha}.$$

### Zero inflation models

A related formulation is the *zero inflation model*, which is a type of *latent class model*. The model accommodates a situation in which the zero outcome can arise in either of two mechanisms. In one regime, the outcome is always zero; in the other, the outcome is generated

by the Poisson or negative binomial process that might also produce a zero. The example suggested in Lambert's (1992) pioneering application is a manufacturing process that produces number of defective parts,  $y_i$ , equal to zero if the process is under control or  $y_i$  equal to a Poisson outcome if the process is not under control. The applicable distribution is

$$\begin{aligned}\text{Prob}(y_i = 0 | \mathbf{x}_i, \mathbf{z}_i) &= \text{Prob}(\text{regime } 0 | \mathbf{z}_i) + \text{Prob}(\text{regime } 1 | \mathbf{z}_i) \text{Prob}(y_i = 0 | \text{regime } 1, \mathbf{x}_i) \\ &= F(r_i | \mathbf{z}_i) + [1 - F(r_i | \mathbf{z}_i)] \text{Prob}(y_i = 0 | \mathbf{x}_i)\end{aligned}$$

$$\text{Prob}(y_i = j | y_i > 0, \mathbf{x}_i, \mathbf{z}_i) = [1 - F(r_i | \mathbf{z}_i)] \text{Prob}(y_i = j | \mathbf{x}_i)$$

The density governing the count process may be the Poisson or negative binomial model. The regime process is typically specified as a logit model, though the probit model is often used as well. Finally, two forms are used for the regime model, the standard probit or logit model with covariate vector,  $\mathbf{z}_i$ , and the zip- $\tau$  form, which takes the form (for logit – Poisson model),

$$\text{Prob}(y_i = 0 | \mathbf{x}_i) = \Lambda(\tau \mathbf{x}_i' \boldsymbol{\beta}) + [1 - \Lambda(\tau \mathbf{x}_i' \boldsymbol{\beta})] \exp(-\lambda_i)$$

$$\text{Prob}(y_i = j | y_i > 0, \mathbf{x}_i) = [1 - \Lambda(\tau \mathbf{x}_i' \boldsymbol{\beta})] \exp(\lambda_i) \lambda_i^j / j!$$

where  $\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$  and  $\tau$  is a single new, free parameter to be estimated. (Researchers usually find that the  $\tau$  form of the model is more restrictive than desired.) The conditional mean function is

$$E[y_i | \mathbf{x}_i, \mathbf{z}_i] = [1 - F(r_i | \mathbf{z}_i)] \lambda_i$$

### *Sample selection*

We consider an extension of the classic model of sample selection [Heckman (1979)] to the models for count outcomes. In the context of the applications considered here, for example, we might consider a sample based on only those individual who have health insurance. The generic model will take the form

$$s_i^* = \mathbf{z}_i' \boldsymbol{\alpha} + u_i \quad u_i \sim N[0,1],$$

$$s_i = \mathbf{1}(s_i^* > 0) \quad (\text{probit selection equation})$$

$$\lambda_i | \varepsilon_i = \exp(\boldsymbol{\beta}' \mathbf{x}_i + \sigma \varepsilon_i) \quad \varepsilon_i \sim N[0,1] \quad (\text{index function with heterogeneity})$$

$$y_i | \mathbf{x}_i, \varepsilon_i \sim \text{Poisson}(y_i | \mathbf{x}_i, \varepsilon_i) \quad (\text{Poisson model for outcome})$$

$$[u_i, \varepsilon_i] \sim N[(0,1), (1, \rho, 1)]$$

$$y_i, \mathbf{x}_i \text{ are observed only when } s_i = 1.$$

The count model is the heterogeneity model suggested earlier with lognormal rather than log gamma heterogeneity. The conventional approach of fitting the probit selection equation, computing an inverse Mills ratio and adding it as an extra regressor in the Poisson model, is inappropriate here. [See Greene (1995, 1997, 2007c).] A formal approach for this model is developed in Terza (1998) and (Greene, 1994, 2006, 2007b,c). Formal results collected in Greene (2006). The generic result for the count model (which can be adapted to the negative binomial or other models) is

$$f(y_i, s_i | \mathbf{x}_i, \mathbf{z}_i) = \int_{-\infty}^{\infty} [(1 - s_i) + s_i f(y_i | \mathbf{x}_i, \varepsilon_i)] \Phi\left((2s_i - 1)[\mathbf{z}'_i \boldsymbol{\alpha}_i + \rho \varepsilon_i] / \sqrt{1 - \rho^2}\right) \phi(\varepsilon_i) d\varepsilon_i,$$

with

$$f(y_i | \mathbf{x}_i, \varepsilon_i) = \frac{\exp(-\lambda_i | \mathbf{x}_i, \varepsilon_i) (\lambda_i | \mathbf{x}_i, \varepsilon_i)^{y_i}}{\Gamma(y_i + 1)}, \quad \lambda_i | \mathbf{x}_i, \varepsilon_i = \exp(\boldsymbol{\beta}' \mathbf{x}_i + \sigma \varepsilon_i).$$

The integral does not exist in closed form, however, the model can be fit by approximating the integrals with Hermite quadrature,

$$\log L_Q = \sum_{i=1}^N \log \left[ \frac{1}{\sqrt{\pi}} \sum_{h=1}^H \omega_h [(1 - s_i) + s_i f(y_i | \mathbf{x}_i, v_h)] \Phi[(2s_i - 1)(\mathbf{z}'_i \boldsymbol{\gamma}_i + \tau v_h)] \right]$$

or simulation, for which the simulated log likelihood is

$$\log L_S = \sum_{i=1}^N \log \frac{1}{R} \sum_{r=1}^R [(1 - s_i) + s_i f(y_i | \mathbf{x}_i, \sigma \varepsilon_{ir})] \Phi[(2s_i - 1)(\mathbf{z}'_i \boldsymbol{\gamma}_i + \tau \varepsilon_{ir})]$$

where  $\gamma = \alpha / (1 - \rho^2)^{1/2}$  and  $\tau = \rho / (1 - \rho^2)^{1/2}$ . There is a minor extension of this model that might be interesting for the health care application examined in this study. The count variables and all the covariates in both equations would be observed for all observations. Thus, to use the full sample of data, the appropriate log likelihood would be

$$f(y_i, z_i | \mathbf{x}_i, \mathbf{z}_i) = \int_{-\infty}^{\infty} f(y_i | \mathbf{x}_i, \varepsilon_i) \Phi((2s_i - 1)[\mathbf{z}'_i \boldsymbol{\gamma}_i + \tau \varepsilon_i]) \phi(\varepsilon_i) d\varepsilon_i,$$

### *Bivariate Poisson model*

The application from which our examples are drawn was a study of the two count variables, *DocVis* and *HospVis*. The authors were interested in a bivariate count model for the two outcomes. One approach to formulating a two equation Poisson is to treat the correlation as arising from the intervention of a latent common Poisson process. The model is

$$\begin{aligned} y_1 &= y_1^* + U \\ y_2 &= y_2^* + U \end{aligned}$$

where  $y_1^*$ ,  $y_2^*$  and  $U$  have three independent Poisson processes. This model is analogous to the seemingly unrelated regressions model. [See King (1989).] The major shortcoming of this approach is that it forces the two variables to be positively correlated. For the application considered here, it is at least possible that the preventive motivation for physician visits could result in a negative correlation between physician and in-patient hospital visits. The approach proposed by Riphahn et al, adapted for a random effects panel data model is  $y_{it,j} \sim \text{Poisson}(\lambda_{it,j})$  where

$$\lambda_{it,j} = \exp(\mathbf{x}_{it,j}' \boldsymbol{\beta} + u_{i,j} + \varepsilon_{it,j}), j = 1, 2.^{26}$$

where the unique heterogeneity  $(\varepsilon_{it,1}, \varepsilon_{it,2})$ , has bivariate normal distribution with correlation  $\rho$  and the random effects, which are constant through time, independent normal distributions. Thus, the

<sup>26</sup> A similar model was estimated by Munkin and Trivedi (1999).

correlation between the conditional means is that induced by the two lognormal variables  $\exp(\varepsilon_{it,1})$  and  $\exp(\varepsilon_{it,2})$ . The implied correlation between  $y_{it,1}$  and  $y_{it,2}$  was not derived. This would differ from  $\rho$ , since both variables have additional variation around the correlated conditional mean functions. The precise result (suppressing the independent variables for convenience) is

$$\text{Cov}[y_1, y_2] = E[\text{Cov}(y_1, y_2 | \varepsilon_1, \varepsilon_2, u)] + \text{Cov}[E[y_1 | \varepsilon_1, u], E[y_2 | \varepsilon_2, u]].$$

The parameter  $\rho$  in this model is merely  $\text{Cov}\{\ln E[y_1 | \varepsilon_1, u], \ln E[y_2 | \varepsilon_2, u]\}$ . How this relates to the unconditional covariance,  $\text{Cov}[y_1, y_2]$  that motivates the analysis remains to be derived.

In order to formulate the log likelihood function, the random components must be integrated out. There are no closed forms for the integrals based on the normal distributions – the problem is similar to that in the sample selection model. The authors used a quadrature procedure to approximate the integrals. The log likelihood could also be maximized by using simulation. Separate models were fit for men and women in the sample. The pooling hypothesis was rejected for all specifications considered.

### 0.6.3 Panel data models

The maximum likelihood estimator of the fixed effects Poisson regression model,

$$\text{Prob}(y_{it} = j | \mathbf{x}_{it}) = \frac{\exp(-\lambda_{it}) \lambda_{it}^j}{j!}, j = 0, 1, \dots; \lambda_{it} = \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta}),$$

is one of a very small number of cases in which the unconditional maximum likelihood estimator is equal to the conditional MLE (conditioned on the sum of the outcomes) and, hence, in which there is no incidental parameters problem. The unconditional MLE of  $\boldsymbol{\beta}$  is consistent in  $n$ . A conditional estimator that is not a function of the fixed effects is found by obtaining the joint distribution of  $(y_{i1}, \dots, y_{iT_i})$  conditional on their sum. For the Poisson model, the conditional probability is:

$$P\left(y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \sum_{t=1}^{T_i} y_{it}\right) = \frac{\left(\sum_{t=1}^{T_i} y_{it}\right)!}{\left(\prod_{t=1}^{T_i} y_{it}!\right)} \prod_{t=1}^{T_i} p_{it}^{y_{it}},$$

where

$$p_{it} = \frac{\exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})}{\sum_{t=1}^{T_i} \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_{it} \boldsymbol{\beta})}{\sum_{t=1}^{T_i} \exp(\mathbf{x}'_{it} \boldsymbol{\beta})}.$$

The contribution of group  $i$  to the conditional log-likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} y_{it} \ln p_{it}.$$

The contribution to  $\ln L$  of a group in which  $y_{it} = 0$  in every period is zero. Such groups fall out of the estimator.

The first order conditions for maximizing the log likelihood function for the Poisson model with respect to the constant terms is

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^T (y_{it} - e^{\alpha_i} \mu_{it}) = 0 \quad \text{where } \mu_{it} = \exp(\mathbf{x}'_{it} \boldsymbol{\beta}).$$

This implies an explicit solution for  $\alpha_i$  in terms of  $\boldsymbol{\beta}$  in this model,

$$\hat{\alpha}_i = \ln \left( \frac{(1/n) \sum_{t=1}^T y_{it}}{(1/n) \sum_{t=1}^T \hat{\mu}_{it}} \right) = \ln \left( \frac{\bar{y}_i}{\bar{\hat{\mu}}_i} \right)$$

[The estimators of  $\alpha_i$  are still inconsistent, however, because their variances are  $O(1/T)$ .]

Unlike the regression or the probit model, this does not require that there be within group variation in  $y_{it}$  - all the values can be the same. It does require that at least one observation for individual  $i$  be nonzero, however.

The random effects probit model is assembled in precisely the fashion of the negative binomial model. If the Poisson model is formed with log gamma heterogeneity, exactly as done earlier, with

$$\ln \lambda_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \ln h_i$$

where  $h_i \sim \text{Gamma}(\theta, \theta)$ ,

then, the conditional joint probability is

$$p(y_{i1}, \dots, y_{iT_i} | h_i) = \prod_{t=1}^{T_i} p(y_{it} | h_i).$$

Then the random effect is swept out by obtaining

$$\begin{aligned} p(y_{i1}, \dots, y_{iT_i}) &= \int_u p(y_{i1}, \dots, y_{iT_i}, h_i) dh_i \\ &= \int_u p(y_{i1}, \dots, y_{iT_i} | h_i) g(h_i) dh_i \\ &= E_h [p(y_{i1}, \dots, y_{iT_i} | h_i)]. \end{aligned}$$

This is exactly the approach used earlier to condition the heterogeneity out of the Poisson model to produce the negative binomial model. The steps produce the negative binomial distribution,

$$p(y_{i1}, \dots, y_{iT_i}) = \frac{[\prod_{t=1}^{T_i} \lambda_{it}^{y_{it}}] \Gamma(\theta + \sum_{t=1}^{T_i} y_{it})}{[\Gamma(\theta) \prod_{t=1}^{T_i} y_{it}!][(\sum_{t=1}^{T_i} \lambda_{it})^{\sum_{t=1}^{T_i} y_{it}}]} Q_i^\theta (1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}},$$

where

$$Q_i = \frac{\theta}{\theta + \sum_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for  $Y_i = \sum_t y_{it}$  with mean  $\Lambda_i = \sum_t \lambda_{it}$ . Thus, the

Hausman, Hall and Griliches (1984) (HHG) report the following conditional density for the fixed effects negative binomial (FENB) model:

$$p\left(y_{i1}, y_{i2}, \dots, y_{iT_i} \mid \sum_{t=1}^{T_i} y_{it}\right) = \frac{\Gamma(1 + \sum_{t=1}^{T_i} y_{it}) \Gamma(\sum_{t=1}^{T_i} \lambda_{it})}{\Gamma(\sum_{t=1}^{T_i} y_{it} + \sum_{t=1}^{T_i} \lambda_{it})} \prod_{t=1}^{T_i} \frac{\Gamma(y_{it} + \lambda_{it})}{\Gamma(1 + y_{it}) \Gamma(\lambda_{it})}$$

which is free of the fixed effects. This is the default FENB formulation used in popular software packages such as SAS, Stata and LIMDEP. Researchers accustomed to the admonishments that fixed effects models cannot contain overall constants or time invariant covariates are sometimes surprised to find (perhaps accidentally) that this fixed effects model allows both. [This issue is explored at length in Allison (2000) and Allison and Waterman (2002).] The resolution of this apparent contradiction is that the HHG FENB model is not obtained by shifting the conditional mean function by the fixed effect,  $\ln \lambda_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i$ , as it is in the Poisson model. Rather, the HHG model is obtained by building the fixed effect into the model as an individual specific  $\theta_i$  in the Negbin 1 form. In the negative binomial models, the conditional mean functions are

$$\text{NB1: } E[y_{it} | \mathbf{x}_{it}] = \theta_i \phi_{it} = \theta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \ln \theta_i),$$

$$\text{NB2: } E[y_{it} | \mathbf{x}_{it}] = \exp(\alpha_i) \phi_{it} = \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i),$$

so, superficially, the formulations do produce the same interpretation. However, the parameter  $\theta_i$  in the NB1 model enters the variance function in a different manner;

$$\text{NB1: } \text{Var} [y_{it} | \mathbf{x}_{it}] = \theta_i \phi_{it} [1 + \theta_i],$$

$$\text{NB2: } \text{Var} [y_{it} | \mathbf{x}_{it}] = \lambda_{it} [1 + \theta \lambda_{it}],$$

The relationship between the mean and the variance is different for the two models. For estimation purposes, one can explain the apparent contradiction noted earlier by observing that in the NB1 formulation, the individual effect is identified separately from the mean in the scedastic (scaling) function. This is not true for the FENB2 form. In order to obtain a counterpart to the HHG model, we would replace  $\theta$  with  $\theta_i$  (and  $\lambda_i$  with  $\lambda_{it}$ ). Greene (2007a) analyzes the more familiar, FENB2 form with the same treatment of  $\lambda_{it}$ . Estimates for both models appear below. Comparison of the suggested NB2 model to the HHG model remains for future investigation.

Once again, theory does not provide a reason to prefer the NB1 formulation over the more familiar NB2 model. The NB1 form does extend beyond the interpretation of the fixed effect as carrying only the sum of all the time invariant effects in the conditional mean function. The appearance of  $\ln \theta_i$  in the conditional mean is an artifact of the exponential mean form;  $\theta_i$  is a scaling parameter in this model. In its favor, the HHG model, being conditionally independent of the fixed effects, finesses the incidental parameters problem – the estimator of  $\boldsymbol{\beta}$  in this model is consistent. This is not the case for the FENB2 form.

Like the fixed effects model, introducing random effects into the negative binomial model adds some additional complexity. We do note, since the negative binomial model derives from the Poisson model by adding latent heterogeneity to the conditional mean, adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time. However, one might prefer to interpret the negative binomial as the density for  $y_{it}$  in its own right, and treat the common effects in the familiar fashion. HHG's (1984) random effects negative binomial model is a hierarchical model that is constructed as follows. The heterogeneity is assumed to enter  $\lambda_{it}$  additively with a gamma distribution with mean 1,  $\Gamma(\theta_i, \theta_i)$ . Then,  $\theta_i/(1+\theta_i)$  is assumed to have a beta distribution with parameters  $a$  and  $b$ . The resulting unconditional density after the heterogeneity is integrated out is

$$p(y_{i1}, y_{i2}, \dots, y_{iT_i}) = \frac{\Gamma(a+b)\Gamma\left(a + \sum_{t=1}^{T_i} \lambda_{it}\right)\Gamma\left(b + \sum_{t=1}^{T_i} y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a + \sum_{t=1}^{T_i} \lambda_{it} + b + \sum_{t=1}^{T_i} y_{it}\right)}.$$

As before, the relationship between the heterogeneity and the conditional mean function is unclear, since the random effect impacts the parameter of the scedastic function. An alternative approach that maintains the essential flavor of the Poisson model (and other random effects models) is to augment the NB2 form with the random effect,

$$\begin{aligned}\text{Prob}(Y = y_{it} | \mathbf{x}_{it}, \varepsilon_i) &= \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it} + 1)\Gamma(\theta)} r_{it}^{y_{it}} (1 - r_{it})^\theta, \\ \lambda_{it} &= \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_i), \\ r_{it} &= \lambda_{it} / (\theta + \lambda_{it}).\end{aligned}$$

We then estimate the parameters by forming the conditional (on  $\varepsilon_i$ ) log likelihood and integrating  $\varepsilon_i$  out either by quadrature or simulation. The parameters are simpler to interpret by this construction. Estimates of the two forms of the random effects model are presented below for a comparison.

#### 0.6.4 Applications

The study by Riphahn et al. (2003) that provided the data we have used in numerous earlier examples analyzed the two count variables *DocVis* (visits to the doctor) and *HospVis* (visits to the hospital). The authors were interested in the joint determination of these two count variables. One of the issues considered in the study was whether the data contained evidence of moral hazard, that is, whether health care utilization as measured by these two outcomes was influenced by the subscription to health insurance. The data contain indicators of two levels of insurance coverage, *Public*, which is the main source of insurance, and *Addon*, which is a secondary optional insurance. In the sample of 27,326 observations (family/years), 24,203 individuals held the public insurance. (There is quite a lot of within group variation in this. Individuals did not routinely obtain the insurance for all periods). Of these 24,203, 23,689 had only public insurance and 514 had both types. (One could not have only the addon insurance.) To explore the issue, we have analyzed the *DocVis* variable with the count data models described above. Figure 1 below shows a histogram for this count variable. (There is a very long tail of extreme observations in these data, extending up to 121.. The histogram omits the 91 observations with *DocVis* greater than 40. All observations are included in sample used to estimate the models.) The exogenous variables in our model are

$$\mathbf{x}_{it} = (1, \text{Age}, \text{Education}, \text{Income}, \text{Kids}, \text{Public})$$

(Variables are described in Table 2. Those listed are a small subset of those used in the original study, chosen here only for a convenient example.)

Table 6 presents the estimates of the several count models. In all specifications, the coefficient on *Public* is positive, large, and highly statistically significant, which is consistent with the results in the authors' study. The large spike at zero in the histogram casts some doubt on the Poisson specification. As a first step in extending the model, we estimated an alternative model that has a distribution that appears more like that in the figure, a *geometric regression model*,

$$\text{Prob}(y_i = j | \mathbf{x}_i) = \pi_i (1 - \pi_i)^j, \pi_i = 1/(1 + \lambda_i), \lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}), j = 0, 1, \dots$$

This is the distribution for the number of failures before the first success in independent trials with success probability equal to  $\pi_i$ . It is suggested here simply as an alternative functional form for the model. The two models are similarly parameterized. The geometric model also has

conditional mean equal to  $(1-\pi_i)/\pi_i = \lambda_i$ , like the Poisson. The variance is equal to  $(1/\pi_i)\lambda_i > \lambda_i$ , so the geometric distribution is overdispersed – it allocates more mass to the zero outcome. Based on the log likelihoods, the Poisson model would be overwhelmingly rejected. However, since the models are not nested, this is not a valid test. Using, instead, the Vuong statistic based on  $v_i = \ln L_i(\text{geometric}) - \ln L_i(\text{Poisson})$ , we obtain a value of +37.89, which, as expected, strongly rejects the Poisson model.

The various formal test statistics strongly reject the hypothesis of equidispersion. Cameron and Trivedi's (1990) semiparametric tests from the Poisson model have  $t$  statistics of 22.147 for  $g_i = \mu_i$  and 22.504 for  $g_i = \mu_i^2$ . Both of these are far larger than the critical value of 1.96. The LM statistic [see Winkelmann (2003)] is 972,714.48, which is also larger than the (any) critical value. On any of these bases, we would reject the hypothesis of equidispersion. The Wald and likelihood ratio tests based on the negative binomial models produce the same conclusion. For comparing the different negative binomial models, note that Negbin 2 is the worst of the three by the likelihood function, though NB1 and NB2 are not directly comparable. On the other hand, in the NBP model, the estimate of  $P$  is more than 10 standard errors from 1.0000 or 2.000, so both NB1 and NB2 would be rejected in favor of the unrestricted NBP form of the model. The NBP and the heterogeneous NB2 model are not nested either, but comparing the log likelihoods, it does appear that the heterogeneous model is substantially superior. We computed the Vuong statistic based on the individual contributions to the log likelihoods, with  $v_i = \ln L_i(\text{NBP}) - \ln L_i(\text{NB2-H})$ . The value of the statistic is -3.27. On this basis, we would reject NBP in favor of NB2-H. Finally, with regard to the original question, the coefficient on *Public* is larger than 10 times the estimated standard error in every specification. We would conclude that the results are consistent with the proposition that there is evidence of moral hazard.

Estimates of the two 2 part models, zero inflated and hurdle, are presented in Table 7. The regime equation for both is assumed to be a logit binary choice model with

$$\mathbf{z}_{it} = (1, \text{Age}, \text{Female}, \text{Married}, \text{Kids}, \text{Income}, \text{Self Employed})$$

There is little theoretical basis for choosing between the two models. The interpretation of the data generating process is quite similar. Each posits a regime in which the individual chooses whether or not to 'participate' in the health care system and a process that generates the count when they do. Nonetheless, there is little doubt that both are improvements on the Poisson regression. The average predicted probability of the zero outcome is 0.04826, so the model predicts  $n\hat{P}_0 = 1,319$  zero observations. The frequency in the sample is 10,135. The counterparts for the ZIP model are 0.36340 and 9,930. The Poisson model is not nested in the ZIP model – setting the ZIP coefficients to zero forces the regime probability to  $\frac{1}{2}$ , not to 1.0. Thus, the models cannot be compared based on their log likelihoods. The Vuong statistic strongly supports the zero inflation model, with  $V = +47.05$ . Similar results are obtained for the hurdle model with the same specification.

The German health care panel data set contains 7,293 individuals with group sizes ranging from 1 to 7. Table 8 presents the fixed and random effects estimates of the equation for *DocVis*. The pooled estimates are also shown for comparison. Overall, the panel data treatments bring large changes in the estimates compared to the pooled estimates. There is also a considerable amount of variation across the specifications. With respect to the parameter of interest, *Public*, we find that the size of the coefficient falls substantially with all panel data treatments. Whether using the pooled, fixed or random effects specifications, the test statistics (Wald, LR) all reject the Poisson model in favor of the negative binomial. Similarly, either common effects specification is preferred to the pooled estimator. There is no simple basis for choosing between the fixed and random effects models, and we have further blurred the distinction by suggesting two formulations of each of them. We do note, the two random effects

estimators are producing similar results, which one might hope for. But, the two fixed effects estimators are producing very different estimates. The NB1 estimates include two coefficients, *Income* and *Education* that are positive, but negative in every other case. Moreover, the coefficient on *Public* which is large and significant throughout the table has become small and less so with the fixed effects estimators.

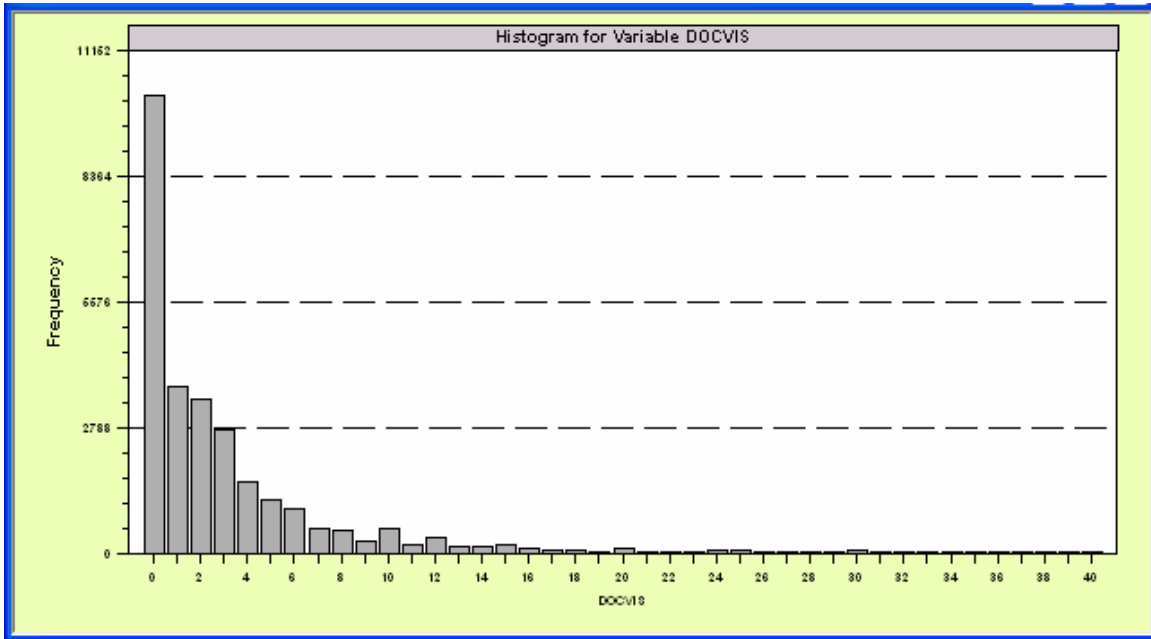


Figure 1 Histogram of count variable *DocVis*

Table 6 Estimated Pooled Models for DocVis (Standard errors in parentheses)

<i>Variable</i>	<i>Poisson</i>	<i>Geometric</i>	<i>Negbin 2</i>	<i>Negbin 2 Heterogeneous</i>	<i>Negbin 1</i>	<i>Negbin P</i>
<b>Constant</b>	0.7162 (0.03287)	0.7579 (0.06314)	0.7628 (0.07247)	0.7928 (0.07459)	0.6848 (0.06807)	0.6517 (0.07759)
<b>Age</b>	0.01844 (0.000332)	0.01809 (0.00669)	0.01803 (0.000792)	0.01704 (0.000815)	0.01585 (0.00070)	0.01907 (0.0008078)
<b>Education</b>	-0.03429 (0.00180)	-0.03799 (0.00343)	-0.03839 (0.003965)	-0.03581 (0.004034)	-0.02381 (0.00370)	-0.03388 (0.004308)
<b>Income</b>	-0.4751 (0.02198)	-0.4278 (0.04137)	-0.4206 (0.04700)	-0.4108 (0.04752)	-0.1892 (0.04452)	-0.3337 (0.05161)
<b>Kids</b>	-0.1582 (0.00796)	-0.1520 (0.01561)	-0.1513 (0.01738)	-0.1568 (0.01773)	-0.1342 (0.01647)	-0.1622 (0.01856)
<b>Public</b>	0.2364 (0.0133)	0.2327 (0.02443)	0.2324 (0.02900)	0.2411 (0.03006)	0.1616 (0.02678)	0.2195 (0.03155)
<b>P</b>	0.0000 (0.0000)	0.0000 (0.0000)	2.0000 (0.0000)	2.0000 (0.0000)	1.0000 (0.0000)	1.5473 (0.03444)
<b>θ</b>	0.0000 (0.0000)	0.0000 (0.0000)	1.9242 (0.02008)	2.6060 (0.05954)	6.1865 (0.06861)	3.2470 (0.1346)
<b>δ (Female)</b>	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	-0.3838 (0.02046)	0.0000 (0.0000)	0.0000 (0.0000)
<b>δ (Married)</b>	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	-0.1359 (0.02307)	0.0000 (0.0000)	0.0000 (0.0000)
<b>Ln L</b>	-104440.3	-61873.55	-60265.49	-60121.77	-60260.68	-60197.15

Table 7 Two part models for DocVis

<b>Variable</b>	<b>Poisson</b>	<b>Poisson/Logit Zero Inflation</b>		<b>Poisson/Logit Hurdle</b>	
	<b>Count</b>	<b>Count</b>	<b>Regime</b>	<b>Count</b>	<b>Regime</b>
<b>Constant</b>	0.7162 (0.03287)	1.3689 (0.01338)	0.4789 (0.0651)	1.4187 (0.0128)	-0.5105 (0.0637)
<b>Age</b>	0.01844 (0.000332)	0.01067 (0.00013)	-0.01984 (0.00133)	0.01059 (0.00012)	0.02068 (0.00131)
<b>Education</b>	-0.03429 (0.00180)	-0.02038 (0.00075)	0.0000 (0.0000)	-0.02215 (0.00072)	0.0000 (0.0000)
<b>Income</b>	-0.4751 (0.02198)	-0.4131 (0.00869)	0.1663 (0.0758)	-0.4560 (0.00831)	-0.2499 (0.0724)
<b>Kids</b>	-0.1582 (0.00796)	-0.08639 (0.00316)	0.2306 (0.0303)	-0.08862 (0.00297)	-0.2378 (0.0297)
<b>Public</b>	0.2364 (0.0133)	0.1573 (0.00604)	0.0000 (0.0000)	0.1547 (0.006037)	0.0000 (0.0000)
<b>Female</b>	0.0000 (0.0000)	0.0000 (0.0000)	-0.58789 (0.0265)	0.0000 (0.0000)	0.5812 (0.0260)
<b>Married</b>	0.0000 (0.0000)	0.0000 (0.0000)	-0.1257 (0.0342)	0.0000 (0.0000)	0.1271 (0.0336)
<b>Self Employed</b>	0.0000 (0.0000)	0.0000 (0.0000)	0.4172 (0.0521)	0.0000 (0.0000)	-0.4137 (0.0513)
<b>log likelihood</b>	-104440.3	-83648.75		-83988.80	

Table 8 Estimated Panel Data Models for Doctor Visits (Standard errors in parentheses)

Variable	Poisson			Negative Binomial				
	Pooled (Robust S.E.)	Fixed Effects	Random Effects	Pooled NB2	Fixed Effects		Random Effects	
					FE-NB1	FE-NB2	HHG-Gamma	Normal
<b>Constant</b>	0.7162 (0.1319)	0.0000	0.4957 (0.05463)	0.7628 (0.07247)	-1.2354 (0.1079)	0.0000	-0.6343 (0.07328)	0.1169 (0.06612)
<b>Age</b>	0.01844 (0.001336)	0.03115 (0.001443)	0.02329 (0.0004458)	0.01803 (0.0007916)	0.02389 (0.001188)	0.04479 (0.002769)	0.01899 (0.0007820)	0.02231 (0.0006969)
<b>Educ</b>	-0.03429 (0.007255)	-0.03803 (0.01733)	-0.03427 (0.004352)	-0.03839 (0.003965)	0.01652 (0.006501)	-0.04589 (0.02967)	-0.01779 (0.004056)	-0.03773 (0.003595)
<b>Income</b>	-0.4751 (.08212)	-0.3030 (0.04104)	-0.2646 (0.01520)	-0.4206 (0.04700)	0.02373 (0.05530)	-0.1968 (0.07320)	-0.08126 (0.04565)	-0.1743 (0.04273)
<b>Kids</b>	-0.1582 (0.03115)	-0.001927 (0.01546)	-0.03854 (0.005272)	-0.1513 (0.01738)	-0.03381 (0.02116)	-0.001274 (0.02920)	-0.1103 (0.01675)	-0.1187 (0.01582)
<b>Public</b>	0.2365 (0.04307)	0.1015 (0.02980)	0.1535 (0.01268)	0.2324 (0.02900)	0.05837 (0.03896)	0.09700 (0.05334)	0.1486 (0.02834)	0.1940 (0.02574)
<b><math>\theta</math></b>	0.0000	0.0000	1.1646 (0.01940)	1.9242 (0.02008)	0.0000	1.9199 (0.02994)	0.0000	1.0808 (0.01203)
<b>a</b>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.1463 (0.05955)	0.0000
<b>b</b>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.8011 (0.1145)	0.0000
<b><math>\sigma</math></b>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9737 (0.008235)
<b>lnL</b>	-104440.3	-60337.13	-71763.13	-60265.49	-34016.16	-49476.36	-58182.52	-58177.66

## 0.7 Multinomial unordered choices

We now extend the random utility, discrete choice model of Sections 0.2 – 0.4 to a setting in which the individual chooses among multiple alternatives. [See Hensher, Rose and Greene (2005).] The random utility model, as before, is

$$U_{it,j} = \mathbf{x}_{it,j}'\boldsymbol{\beta} + \mathbf{z}_{it}'\boldsymbol{\gamma} + \varepsilon_{it,j}, j = 1, \dots, J_{it}, t = 1, \dots, T_i,$$

where, as before, we consider individual  $i$  in choice situation  $t$ , choosing among a possibly variable number of choices,  $J_{it}$  and a possibly individual specific number of choice situations. For the present, for convenience, we assume  $T_i = 1$  – a single choice situation. This will be generalized later. The extension to variable choice set sizes,  $J_{it}$ , turns out to be essentially a minor modification of the mathematics, so it will also prove convenient to assume  $J_{it}$  is fixed at  $J$ . The random utility model is, thus,

$$U_{i,j} = \mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{i,j}, j = 1, \dots, J, i = 1, \dots, n.$$

The earlier assumptions are extended as well. The axioms of choice will imply that preferences are transitive, reflexive, and complete. Thus, in any choice situation, the individual will make a choice, that that choice,  $j_i$  will be such that

$$U_{i,j_i} > U_{i,m} \text{ for all } m = 1, \dots, J \text{ and } m \neq j_i.$$

Reverting back to the classical problem of utility maximization over continuous choices subject to a budget constraint produces the complete set of demands,  $\mathbf{d}_i(\text{prices, income})$ . Inserting the demands back into the utility function produces the indirect utility function,

$$U_i^* = U_i[\mathbf{d}, \mathbf{x}(\text{prices, income})].$$

This formulation is convenient for discrete choice modeling, as the data typically observed on the right hand sides of the model equations will be income, prices, and other characteristics of the individual such as age and sex, and attributes of the choices, such as model or type. The random utility model for multinomial unordered choices, is then taken to be defined over the indirect utilities.

### 0.7.1 Multinomial logit and multinomial probit models

Not all stochastic specifications for  $\varepsilon_{i,j}$  are consistent with utility maximization. McFadden (1981) showed that the i.i.d., type 1 extreme value distribution,

$$F(\varepsilon_{i,j}) = \exp(-\exp(-\varepsilon_{i,j})), j = 1, \dots, J, i = 1, \dots, n,$$

produces a probabilistic choice model that is consistent with utility maximization. The resulting choice probabilities are

$$\text{Prob}(d_{i,j} = 1 \mid \mathbf{X}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{x}_{i,j}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_j)}{\sum_{m=1}^J \exp(\mathbf{x}_{i,m}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_m)}, d_{i,j} = 1 \text{ if } U_{i,j_i} > U_{i,m}, m = 1, \dots, J \text{ and } m \neq j_i.$$

This is the *multinomial logit model*. The components,  $\mathbf{x}_{i,j}$  are the attributes of the choices (prices, features, etc.) while  $\mathbf{z}_i$  is the characteristics of the individual (income, age, sex). We noted at the outset of Section 0.2 that identification of the model parameters requires that  $\boldsymbol{\gamma}$  vary across the choices. Thus, the full model

$$\text{Prob}(d_{i,j} = 1 \mid \mathbf{X}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{x}'_{i,j} \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma}_j)}{\sum_{m=1}^J \exp(\mathbf{x}'_{i,m} \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma}_m)}, \boldsymbol{\gamma}_J = \mathbf{0},$$

$$d_{i,j} = 1 \text{ if } U_{i,j_i} > U_{i,m}, m = 1, \dots, J \text{ and } m \neq j_i.$$

The log likelihood function is

$$\ln L = \sum_{i=1}^n \sum_{j=1}^J d_{i,j} \ln \left[ \frac{\exp(\mathbf{x}'_{i,j} \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma}_j)}{\sum_{m=1}^J \exp(\mathbf{x}'_{i,m} \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma}_m)} \right]$$

The multinomial logit specification implies the peculiar restriction that

$$\frac{\partial \ln \text{Prob}(\text{choice} = j)}{\partial \mathbf{x}_{i,m}} = [1(j = m) - \text{Prob}(\text{choice} = m)] \boldsymbol{\beta}$$

Thus, the impact of a change in an attribute of a particular choice on the set of choice probabilities is the same for all (other) choices. For example, in our application,

$$\frac{\partial \ln P_{\text{TRAIN}}}{\partial \text{Cost}_{\text{AIR}}} = \frac{\partial \ln P_{\text{BUS}}}{\partial \text{Cost}_{\text{AIR}}} = \frac{\partial \ln P_{\text{CAR}}}{\partial \text{Cost}_{\text{AIR}}} = (-P_{\text{AIR}}) \boldsymbol{\beta}_{\text{Cost}}$$

This striking result, termed the *independence from irrelevant alternatives* (IIA), follows from the initial assumptions of independent and identical distributions for  $\varepsilon_{i,j}$ . This is a major shortcoming of the model, and has motivated much of the research on specification of the discrete choice models. Many model extensions have been proposed, including a heteroscedastic extreme value model [Bhat(1995)] and the DOGIT (dodging the logit model, Gaudry and Dagenais (1979)) and a host of others. The major extensions of the canonical multinomial logit model (MNL) have been the multinomial probit model, the nested logit model and the current frontier, the mixed logit model. We consider each of these in turn.

#### *Multinomial Probit Model*

The multinomial probit (MNP) model [Daganzo (1979)] replaces the i.i.d. assumptions of the multinomial logit model with a multivariate normality assumption,

$$\boldsymbol{\varepsilon}_i \sim N_j[\mathbf{0}, \boldsymbol{\Sigma}]$$

This specification relaxes the independence assumption. In principle, it can also relax the assumption of identical (marginal) distributions as well. Recall that since only the most preferred choice is revealed, information about utilities is obtained in the form of differences,  $U_{i,j} - U_{i,m}$ . It follows that identification restrictions are required – only some, or certain combinations of elements of  $\boldsymbol{\Sigma}$  are estimable. The simplest approach to securing identification that is used in

practice is to impose that the last row of  $\Sigma$  be equal to  $(0,0,\dots,1)$ , and one other diagonal element also equal 1. The remaining elements of  $\Sigma$  may be unrestricted, subject to the requirement that the matrix be positive definite. This can be done by a Cholesky decomposition,  $\Sigma = CC'$  where  $C$  is a lower triangular matrix.

The MNP model relaxes the IIA assumptions. The shortcoming of the model is the computational demands. The relevant probabilities that enter the log likelihood function and its derivatives must be approximated by simulation. The GHK simulator [Lerman and Manski (1977), Geweke et al. (1994)] is commonly used. The Gibbs sampler with noninformative priors [Allenby and Rossi (1999) and Rossi and Allenby (2003)] has also proved useful for estimating the model parameters. Even with the GHK simulator, however, computation of the probabilities by simulation is time consuming.

### 0.7.2 Nested logit models

The nested logit model allows for grouping of alternatives into ‘nests’ with correlation across elements in a group. The natural analogy is to a ‘tree structure.’ For example, Figure 2 suggests an elaborate, three level treatment of an eight alternative choice set::

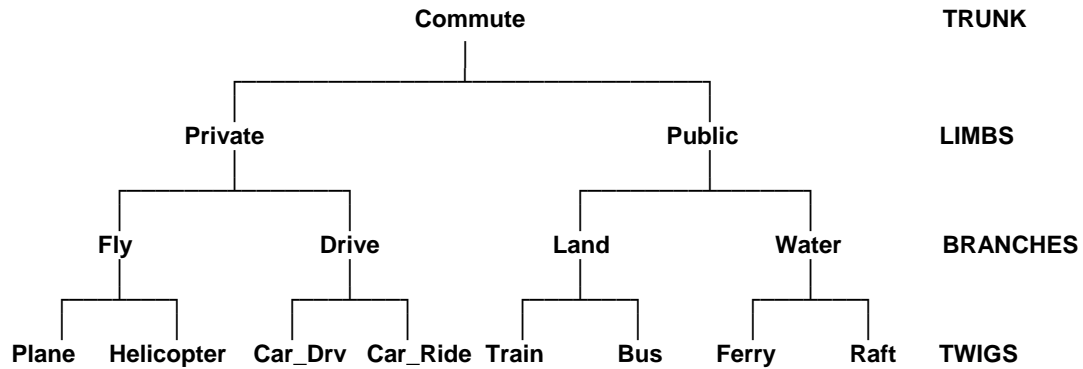


Figure 2 Nested choice set.

The specific choice probabilities are redefined to be the conditional probability of alternative  $j$  in branch  $b$ , limb  $l$ , and trunk  $r$ ,  $j/b,l,r$ . At the next level up the tree, we define the conditional probability of choosing a particular branch in limb  $l$ , trunk  $r$ ,  $b/l,r$ , the conditional probability of choosing limb in trunk  $r$ ,  $l/r$ , and, finally, the probability of choosing trunk  $r$ . By the laws of probability, the unconditional probability of the observed choices made by an individual is

$$P(j,b,l,r) = P(j|b,l,r) \times P(b/l,r) \times P(l/r) \times P(r).$$

This is the contribution of an individual observation to the likelihood function for the sample. (Note in our example, there is only one trunk, so  $P(r) = 1$ .)

The two level nested logit model is the leading case, and occupies most of the received applications. In this instance, a common specification places the individual specific characteristics, such as demographic variables, in the branch probabilities. For this basic model, then,

$$P(j/b) = \frac{\exp(\mathbf{x}'_{j|b}\boldsymbol{\beta})}{\sum_{q|b} \exp(\mathbf{x}'_{q|b}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_{j|b}\boldsymbol{\beta})}{\exp(I V_b)},$$

where  $IV_b$  is the inclusive value for branch  $b$ ,

$$IV_b = \log \sum_{q|b} \exp(\mathbf{x}_{q|b}'\boldsymbol{\beta}).$$

At the next level up the tree, we define the conditional probability of choosing a particular branch

$$P(b) = \frac{\exp[\lambda_b(\mathbf{z}'_i\boldsymbol{\gamma}_b + IV_b)]}{\sum_s \exp[\lambda_s(\mathbf{z}'_i\boldsymbol{\gamma}_s + IV_s)]} = \frac{\exp[\lambda_b(\mathbf{z}'_i\boldsymbol{\gamma}_b + IV_b)]}{\exp(IV)},$$

where  $I$  is the inclusive value for the model,

$$IV = \log \sum_s \exp[\lambda_s(\mathbf{z}'_i\boldsymbol{\delta}_s + IV_s)].$$

The original MNL results if the inclusive parameters,  $\lambda_s$  are all equal to one.

Alternative normalizations and aspects of the nested logit model are discussed in Hensher and Greene (1999) and Hunt (2000). A second form moves the scaling down to the twig level, rather than at the branch level. Here it is made explicit that within a branch, the scaling must be the same for alternatives, but it can differ between the branches.

$$P(j|b) = \frac{\exp[\mu_b(\mathbf{x}'_{j|b}\boldsymbol{\beta})]}{\sum_{q|b} \exp[\mu_b(\mathbf{x}'_{q|b}\boldsymbol{\beta})]} = \frac{\exp[\mu_b(\mathbf{x}'_{j|b}\boldsymbol{\beta})]}{\exp(IV_b)}.$$

Note in the summation in the inclusive value that the scaling parameter is not varying with the summation index. It is the same for all twigs in the branch. Now,  $IV_b$  is the inclusive value for branch  $b$ ,

$$IV_b = \log \sum_{q|b} \exp[\mu_b(\mathbf{x}_{q|b}'\boldsymbol{\beta})].$$

At the next level up the tree, we define the conditional probability of choosing the particular branch,

$$P(b) = \frac{\exp[(\mathbf{z}'_i\boldsymbol{\gamma}_b + (1/\mu_b)IV_b)]}{\sum_s \exp[(\mathbf{z}'_i\boldsymbol{\gamma}_s + (1/\mu_s)IV_s)]} = \frac{\exp[(\mathbf{z}'_i\boldsymbol{\gamma}_b + (1/\mu_b)IV_b)]}{\exp(IV)},$$

where  $IV_l$  is the inclusive value for limb  $l$ ,

$$I_l = \log \sum_{s|l} \exp[\gamma_l(\boldsymbol{\alpha}'\mathbf{y}_{s|l} + (1/\mu_{s|l})IV_{s|l})].$$

In the nested logit model with  $P(j,b,l,r) = P(j|b,l,r) \times P(b|l,r) \times P(l|r) \times P(r)$ , the marginal effect of a change in attribute ' $k$ ' in the utility function for alternative ' $J$ ' in branch ' $B$ ' of limb ' $L$ ' of trunk ' $R$ ' on the probability of choice ' $j$ ' in branch ' $b$ ' of limb ' $l$ ' of trunk ' $r$ ' is computed using the following result: Lower case letters indicate the twig, branch, limb and trunk of the outcome upon which the effect is being exerted. Upper case letters indicate the twig, branch, limb and trunk which contain the outcome whose attribute is being changed:

$$\frac{\partial \log P(\text{alt} = j, \text{limb} = l, \text{branch} = b, \text{trunk} = r)}{\partial x(k) | \text{alt} = J, \text{limb} = L, \text{branch} = B, \text{trunk} = r} = D(k | J, B, L, R) = \Delta(k) \times F,$$

where  $\Delta(k)$  = coefficient on  $x(k)$  in  $U(J/B,L,R)$

$$\begin{aligned} \text{and } F &= \mathbf{1}(r=R) \times \mathbf{1}(l=L) \times \mathbf{1}(b=B) \times [\mathbf{1}(j=J) - P(J/BLR)] && \text{(trunk effect),} \\ &\quad \mathbf{1}(r=R) \times \mathbf{1}(l=L) \times [\mathbf{1}(b=B) - P(B/LR)] \times P(J/BLR) \times \tau_{B/LR} && \text{(limb effect),} \\ &\quad \mathbf{1}(r=R) \times [\mathbf{1}(l=L) - P(L/R)] \times P(B/LR) \times P(J/BLR) \times \tau_{B/LR} \times \sigma_{L/R} && \text{(branch effect),} \\ &\quad [\mathbf{1}(r=R) - P(R)] \times P(L/R) \times P(B/LR) \times P(J/BLR) \times \tau_{B/LR} \times \sigma_{L/R} \times \phi_R && \text{(twig effect).} \end{aligned}$$

where  $\tau_{B/LR}$ ,  $\sigma_{L/R}$  and  $\phi_R$  are parameters in the MNL probabilities. The marginal effect is

$$\partial P(j,b,l,r)/\partial x(k)|_{J,B,L,R} = P(j,b,l,r) \Delta(k) F.$$

A marginal effect has four components, an effect on the probability of the particular trunk, one on the probability for the limb, one for the branch, and one for the probability for the twig. (Note that with one trunk,  $P(l) = P(1) = 1$ , and likewise for limbs and branches.) For continuous variables, such as cost, it is common to report, instead,

$$\text{Elasticity} = x(k)|_{J,B,L,R} \times \Delta(k|_{J,B,L,R}) \times F.$$

The formulation of the nested logit model imposes no restrictions on the inclusive value parameters. However, the assumption of utility maximization and the stochastic underpinnings of the model do imply certain restrictions. For the former, in principle, the inclusive value parameters must be between zero and one. For the latter, the restrictions are implied by the way that the random terms in the utility functions are constructed. In particular, the nesting aspect of the model is obtained by writing

$$\varepsilon_{j|b,l,r} = u_{j|b,l,r} + v_{b|l,r}.$$

That is, within a branch, the random terms are viewed as the sum of a unique component and a common component. This has certain implications for the structure of the scale parameters in the model. In particular, it is the source of the oft cited (and oft violated) constraint that the IV parameters must lie between zero and one. These are explored in Hunt (2000) and Hensher and Greene (1999).

### 0.7.3 Mixed logit and error components models

This model is somewhat similar to the random coefficients model for linear regressions. (See Bhat (1996), Jain, Vilcassim, and Chintagunta (1994), Revelt and Train (1998), Train (2003), and Berry, Levinsohn, and Pakes (1995).) The model formulation is a one level multinomial logit model, for individuals  $i = 1, \dots, n$  in choice setting  $t$ . We begin with the basic form of the multinomial logit model, with alternative specific constants  $\alpha_{ji}$  and attributes  $\mathbf{x}_{ji}$ ,

$$\text{Prob}(y_{it} = j | \mathbf{X}_{it}) = \frac{\exp(\alpha_{ji} + \mathbf{x}'_{it,j} \boldsymbol{\beta}_i)}{\sum_{q=1}^{J_{it}} \exp(\alpha_{qi} + \mathbf{x}'_{it,q} \boldsymbol{\beta}_i)}.$$

The random parameters model emerges as the form of the individual specific parameter vector,  $\boldsymbol{\beta}_i$ ,

is developed. The most familiar, simplest version of the model specifies

$$\begin{aligned}\beta_{ki} &= \beta_k + \sigma_k v_{ki}, \\ \alpha_{ji} &= \alpha_j + \sigma_j v_{ji},\end{aligned}$$

where  $\beta_k$  is the population mean,  $v_{ki}$  is the individual specific heterogeneity, with mean zero and standard deviation one, and  $\sigma_k$  is the standard deviation of the distribution of  $\beta_{ki}$ s around  $\beta_k$ . The term ‘mixed logit’ is often used in the literature [e.g., Revelt and Train (1998) and, especially, McFadden and Train (2000)]. The choice specific constants,  $\alpha_{ji}$  and the elements of  $\beta_i$  are distributed randomly across individuals with fixed means. A refinement of the model is to allow the means of the parameter distributions to be heterogeneous with observed data,  $\mathbf{z}_i$ , (which does not include a constant). This would be a set of choice invariant characteristics that produce individual heterogeneity in the means of the randomly distributed coefficients so that

$$\beta_{ki} = \beta_k + \mathbf{z}_i' \boldsymbol{\delta}_k + \sigma_k v_{ki},$$

and likewise for the constants. The model is not limited to the normal distribution. One important variation is the lognormal model,

$$\beta_{ki} = \exp(\beta_k + \mathbf{z}_i' \boldsymbol{\delta}_k + \sigma_k v_{ki}).$$

The  $v_{ki}$ s are individual and choice specific, unobserved random disturbances - the source of the heterogeneity. Thus, as stated above, in the population, if the random terms are normally distributed,

$$\beta_{ki} \sim \text{Normal or Lognormal} [\beta_k + \mathbf{z}_i' \boldsymbol{\delta}_k, \sigma_k^2].$$

(Other distributions may be specified.) For the full vector of  $K$  random coefficients in the model, we may write the full set of random parameters as

$$\beta_i = \boldsymbol{\beta} + \Delta \mathbf{z}_i + \Gamma \mathbf{v}_i.$$

where  $\Gamma$  is a diagonal matrix which contains  $\sigma_k$  on its diagonal.

Hensher and Greene (2006) and Greene and Hensher (2006) have developed a counterpart to the random effects model that essentially generalizes the mixed logit model to a stochastic form of the nested logit model. The general notation is fairly cumbersome, but an example suffices to develop the model structure. Consider a four outcome choice set, Air, Train, Bus, Car. The utility functions in an MNL or mixed logit model could be

$$\begin{aligned}U_{it,Air} &= \alpha_{Air} + \mathbf{x}_{it,Air}' \boldsymbol{\beta}_i && + \varepsilon_{it,Air} + \theta_1 E_{i,Private} \\ U_{it,Train} &= \alpha_{Train} + \mathbf{x}_{it,Train}' \boldsymbol{\beta}_i && + \varepsilon_{it,Train} && + \theta_2 E_{i,Public} \\ U_{it,Bus} &= \alpha_{Bus} + \mathbf{x}_{it,Bus}' \boldsymbol{\beta}_i && + \varepsilon_{it,Bus} && + \theta_2 E_{i,Public} \\ U_{it,Car} &= && \mathbf{x}_{it,Car}' \boldsymbol{\beta}_i && + \varepsilon_{it,Car} + \theta_1 E_{i,Private}\end{aligned}$$

where the components,  $E_{i,Private}$  and  $E_{i,Public}$  are independent, normally distributed random elements of the utility functions. Thus, this is a two level nested logit model.

The probabilities defined above are conditioned on the random terms,  $\boldsymbol{\varepsilon}_i$  and the error components,  $\mathbf{E}_i$ . The unconditional probabilities are obtained by integrating  $v_{ik}$  and  $E_{im}$  out of the conditional probabilities:  $P_j = E_{\mathbf{v}, \mathbf{E}}[P(j|\boldsymbol{\varepsilon}_i, \mathbf{E}_i)]$ . This is a multiple integral which does not exist in closed form. The integral is approximated by simulation. [See Hensher and Greene (2006) and

Greene (2007) for discussion.] Parameters are estimated by maximizing the simulated log likelihood,

### 0.7.4 Applications

The multinomial choice models are illustrated with a well known data survey of commuters between Sydney and Melbourne [see Greene (2007) and references cited.] A sample of 210 travelers were asked which of four travel modes they chose, among Air, Train, Bus or Car.. The variables used in the models are

TTME = Terminal time, in minutes, zero for car,  
 INVT = In vehicle time for the journey,  
 GC = generalized cost = in vehicle cost + a wage times INVT,  
 HINC = household income,  
 PSIZE = traveling party size.

Descriptive statistics for the data used in estimation are shown in Table 9. We note before beginning, the sample proportions for the four travel modes in this sample are 0.27619, 0.30000, 0.14286 and 0.28095, respectively. Long study of this market revealed that the population values of these proportions would be closer to 0.14, 0.13, 0.09 and 0.64, respectively. The sample observations were deliberately drawn so that the car alternative received fewer observations than random sampling would predict. The sample is *choice based*. A general adjustment for that phenomenon is the Manski-Lerman (1977) WESML correction, which consists of two parts. First, we would fit a weighted log likelihood,

$$\ln L(\text{WESML}) = \sum_{i=1}^n \sum_{j=1}^J \frac{\pi_j}{p_j} d_{ij} \ln \Pi_{ij}$$

where  $d_{ij} = 1$  if individual  $i$  chooses alternative  $j$  and 0 otherwise,  $\pi_j$  is the true population proportion,  $p_j$  is the sample proportion, and  $\Pi_{ij}$  is the probability for outcome  $j$  implied by the model. The second aspect of the correction is to use a sandwich style corrected estimator for the asymptotic covariance matrix of the MLE,

$$V(\text{WESML}) = \mathbf{H}^{-1} (\mathbf{G}'\mathbf{G}) \mathbf{H}^{-1}$$

where  $\mathbf{H}$  is the inverse of the (weighted) Hessian and  $(\mathbf{G}'\mathbf{G})^{-1}$  would be the BHHH estimator based on first derivatives. The results to follow do not include this correction – the results in the example would change slightly if they were incorporated.

We fit a variety of models. The same utility functions were specified for all:

$$\begin{aligned} U_{i,\text{AIR}} &= \alpha_{\text{AIR}} + \beta_{tt} TTME_{i,\text{AIR}} + \beta_{it} INVT_{i,\text{AIR}} + \beta_{gc} GC_{i,\text{AIR}} + \gamma_A HINC_i + \varepsilon_{i,\text{AIR}}, \\ U_{i,\text{TRAIN}} &= \alpha_{\text{TRAIN}} + \beta_{tt} TTME_{i,\text{TRAIN}} + \beta_{it} INVT_{i,\text{TRAIN}} + \beta_{gc} GC_{i,\text{TRAIN}} + \varepsilon_{i,\text{TRAIN}}, \\ U_{i,\text{BUS}} &= \alpha_{\text{BUS}} + \beta_{tt} TTME_{i,\text{BUS}} + \beta_{it} INVT_{i,\text{BUS}} + \beta_{gc} GC_{i,\text{BUS}} + \varepsilon_{i,\text{BUS}}, \\ U_{i,\text{CAR}} &= \beta_{tt} TTME_{i,\text{CAR}} + \beta_{it} INVT_{i,\text{CAR}} + \beta_{gc} GC_{i,\text{CAR}} + \varepsilon_{i,\text{CAR}}, \end{aligned}$$

Model MNL is the base case multinomial logit model. Model MNP is the multinomial probit

model. The three NL models are nested logit models with different tree structure,

- NL(1) = Private(air,car), Public(train,bus).
- NL(2) = Fly(air), Ground(train, bus, car)
- NL(3) = Fly(air), Rail(train), Drive(car), Autobus(bus).

For the third of these, one of the inclusive value parameters,  $\mu_j$  must be constrained to equal one. Model HEV is the extreme value model with the variances allowed to differ across utility functions. In addition, we introduced heteroscedasticity in the model, so that

$$\text{Var}[\varepsilon_{ij}] = \sigma_j^2 \times \exp(\theta \text{ Party Size}_i).$$

Finally, the last model is a random parameters specification in which the parameters on *TTME*, *INVT* and *GC* are allowed to vary randomly across individuals.

Table 9 Descriptive Statistics for Variables

<i>Variable</i>	<i>Mean</i>	<i>Std.Dev.</i>	<i>Mean</i>	<i>Std.Dev.</i>
<i>Air</i>	<i>All 210 Observations</i>		<i>58 Observations that chose AIR</i>	
<i>TTME</i>	61.010	15.719	46.534	24.389
<i>INVT</i>	133.710	48.521	124.828	50.288
<i>GC</i>	102.648	30.575	113.552	33.198
<i>PSIZE</i>	1.743	1.012	1.569	.819
<i>HINC</i>	34.548	19.711	41.724	19.115
<i>Train</i>	<i>All 210 Observations</i>		<i>63 Observations that chose TRAIN</i>	
<i>TTME</i>	35.690	12.279	28.524	19.354
<i>INVT</i>	608.286	251.797	532.667	249.360
<i>GC</i>	130.200	58.235	106.619	49.601
<i>PSIZE</i>	1.743	1.012	1.667	.898
<i>HINC</i>	34.548	19.711	23.063	17.287
<i>Bus</i>	<i>All 210 Observations</i>		<i>30 Observations that chose BUS</i>	
<i>TTME</i>	41.657	12.077	25.200	14.919
<i>INVT</i>	629.462	235.408	618.833	273.610
<i>GC</i>	115.257	44.934	108.133	43.244
<i>PSIZE</i>	1.743	1.012	1.333	.661
<i>HINC</i>	34.548	19.711	29.700	16.851
<i>Car</i>	<i>All 210 Observations</i>		<i>59 Observations that chose CAR</i>	
<i>TTME</i>	.000	.000	.000	.000
<i>INVT</i>	573.205	274.855	527.373	301.131
<i>GC</i>	95.414	46.827	89.085	49.830
<i>PSIZE</i>	1.743	1.012	2.203	1.270
<i>HINC</i>	34.548	19.711	42.220	17.685

There is no useable scalar fit measure for the multinomial choice model. (The Pseudo-R<sup>2</sup> was proposed for this model but, as noted earlier, is not an effective measure of fit.) One approach to assessing model fit is a cross tabulation of actual vs. predicted outcomes. The computation would be

$$\hat{n}_{mj} = \text{int} \left[ \sum_{i=1}^n \hat{P}_{i,m} d_{i,j} \right]$$

where  $\hat{P}_{i,m}$  is the predicted probability for outcome m and  $d_{i,j}$  is the binary indicator for whether individual i chose alternative j. Table 11 reports this computation for the multinomial logit model

Table 10 Estimated Multinomial Choice Models (Standard errors in parentheses)

	<i>MNL</i>	<i>MNP</i>	<i>NL(1)</i>	<i>NL(2)</i>	<i>NL(3)</i>	<i>HEV</i>	<i>RPL</i>
$\alpha_{AIR}$	3.139 (0.984)	-2.769 (1.997)	1.110 (.877)	3.261 (.879)	1.825 (.621)	2.405 (2.692)	6.930 (4.053)
$\alpha_{TRAIN}$	3.558 (0.443)	3.137 (1.0599)	1.468 (.452)	3.039 (.601)	2.113 (.493)	6.701 (2.852)	17.994 (4.745)
$\alpha_{BUS}$	3.134 (0.452)	2.581 (.419)	.971 (.475)	2.721 (.604)	1.877 (.746)	6.150 (2.483)	16.556 (4.585)
$\alpha_{CAR}$	0.000 (0.000)	.000 (0.000)	.000 (0.000)	.000 (0.000)	.000 (0.000)	.000 (0.000)	.000 (0.000)
<i>TermTime</i>	-0.0963 (0.0103)	-.0548 (.0227)	-.0655 (.0116)	-.0742 (.00134)	-.0415 (.0148)	-.164 (.0799)	-.385 (.0857)
<i>Inv.Time</i>	-.00379 (.00118)	-.00447 (.00139)	-.00422 (.000919)	-.0167 (.00142)	-.00767 (.00197)	-.00744 (.00300)	-.0241 (.00589)
<i>Gen.Cost</i>	-.00139 (.00623)	-.0183 (.00827)	-.000449 (.00467)	.00639 (.00679)	-.00051 (.00340)	-.0299 (.0185)	-.0397 (.0238)
<i>Income</i>	.0185 (.0108)	.0702 (.0398)	.0169 (.00691)	.0195 (.00878)	.00868 (.00389)	.0604 (.0456)	.156 (.0715)
<i>Scale(1)</i>		5.073 (2.172)	3.097 (.627)	1.278 (.289)	3.400 (1.238)	.386 (.189)	.261 (.0794)
<i>Scale(2)</i>		1.221 (.911)	1.989 (.423)	.197 (.0679)	1.0839 (.109)	.745 (.376)	.0176 (.00564)
<i>Scale(3)</i>		1.000 (0.000)			1.130 (.144)	.964 (.587)	.0369 (.0350)
<i>Scale(4)</i>		1.000 (0.000)			1.000 (0.000)	1.000 (0.000)	
<i>Party Size</i>						-.208 (.0739)	
$\rho(Air,Train)$		.736 (.323)					
$\rho(Air,Bus)$		.649 (.475)					
$\rho(Train,Bus)$		.655 (.292)					
$\ln L$	-193.498	-191.826	-178.714	-166.366	-190.930	-186.174	-168.109

MNP, scale parameters are standard deviations, NL1, scale parameters are IV parameters  
 NL2, scale parameters are IV parameters, NL(3) Scale parameters are IV parameters  
 HEV, scale parameters are  $\sigma_j$ , RPL, scale parameters are sd's of random parameters

Table 11 Predicted vs. Actual Choices<sup>a</sup>

<i>Actual Outcomes</i>	<i>Predicted Outcomes</i>								
	<i>Air</i>		<i>Train</i>		<i>Bus</i>		<i>Car</i>		<i>Total</i>
	<i>MNL</i>	<i>NL</i>	<i>MNL</i>	<i>NL</i>	<i>MNL</i>	<i>NL</i>	<i>MNL</i>	<i>NL</i>	
<i>Air</i>	33	32	8	7	4	4	13	15	58
<i>Train</i>	7	6	37	49	5	7	14	11	63
<i>Bus</i>	4	3	5	7	16	16	5	4	30
<i>Car</i>	14	16	12	9	5	4	27	31	59
<i>Total</i>	58	57	63	63	30	30	59	61	210

<sup>a</sup> Column totals subject to rounding error. Row totals are actual counts.

and for the first nested logit model in Table 10 (NL(1)). We can see from the counts on the diagonals of the matrix, by this measure, the nested logit model fits slightly better. The improvement in fit is achieved by a much better match of the model predictions for the choice of Train.

Table 12 lists the estimates of the elasticities of the choice probabilities with respect to changes in the generalized cost of each mode. The force of the IID assumptions of the multinomial logit model can be seen in the cross elasticities. For example, the elasticity of the choice probabilities for Train, Bus and Car with respect to changes in GC of Air are all 0.0435. The counterparts for Train, Bus and Car are -.0455, 0,0210 and 0.0346, respectively. None of the other models listed have this property.

Table 12 Estimated Elasticities with Respect to Changes in GC

Effects is on choice of:	CG changes in choices:							
	Air	Train	Bus	Car				
<b>Air</b>	MNL	-.0994	MNL	.0455	MNL	.0210	MNL	.0346
	MNP	-.5230	MNP	.3060	MNP	.1179	MNP	.1006
	NL (2)	.595-	NL (2)	-.0310	NL (2)	-.0200	NL (2)	-.0430
	HEV	-.9158	HEV	.3771	HEV	.2339	HEV	.2144
	RPL	-.4808	RPL	.2361	RPL	.1440	RPL	.0663
<b>Train</b>	MNL	.0435	MNL	-.1357	MNL	.0210	MNL	.0346
	MNP	.3889	MNP	-3.4650	MNP	1.1148	MNP	.9416
	NL (2)	-.2440	NL (2)	-.2160	NL (2)	-.127	NL (2)	.5420
	HEV	.3443	HEV	-1.7389	HEV	.4105	HEV	.4621
	RPL	.3167	RPL	-1.4151	RPL	.5715	RPL	.2360
<b>Bus</b>	MNL	.0435	MNL	.0455	MNL	-.1394	MNL	.0346
	MNP	.2859	MNP	2.454	MNP	-4.4750	MNP	1.2686
	NL (2)	-.2440	NL (2)	-.2160	NL (2)	.6100	NL (2)	-.2900
	HEV	.4744	HEV	1.2723	HEV	-3.1008	HEV	.8358
	RPL	.7109	RPL	1.8434	RPL	-2.9242	RPL	.3246
<b>Car</b>	MNL	.0435	MNL	.0455	MNL	.0210	MNL	-.0982
	MNP	.1113	MNP	.8592	MNP	.5587	MNP	-1.4023
	NL (2)	-.2440	NL (2)	.3940	NL (2)	-.1270	NL (2)	-.2900
	HEV	.4133	HEV	.8108	HEV	.6190	HEV	-1.7829
	RPL	.2489	RPL	.6300	RPL	.2973	RPL	-1.0332

## 0.8 Summary and conclusions

The preceding has outlined the basic modeling frameworks that are used in analyzing microeconomic data when the response variable corresponds to a discrete choice. The essential binary choice model is the foundation for a vast array of applications and theoretical developments. The full set of results for the fully parametric models based on the normal distribution as well as many non- and semiparametric models are well established. Ongoing contemporary theoretical research is largely focused on less parametric approaches and on panel data. The parametric models developed here still overwhelmingly dominate the received applications.

## References

- Abramovitz, M., and I. Stegun. *Handbook of Mathematical Functions*. New York: Dover Press, 1971.
- Abrevaya, J., "The Equivalence of Two Estimators of the Fixed Effects Logit Model," *Economics Letters*, 55, 1, 1997, pp. 41-44.
- Abrevaya, J. and J. Huang, "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, 73, 4, 2005, pp. 1175-1204.
- Akin, J., D. Guilkey and R. Sickles, "A Random Coefficient Probit Model with an Application to a Study of Migration," *Journal of Econometrics*, 11, 1979, pp. 233-246.
- Albert, J., and S. Chib. "Bayesian Analysis of Binary and Polytomous Response Data." *Journal of the American Statistical Association*, 88, 1993a, pp. 669-679.
- Aldrich, J., and F. Nelson. *Linear Probability, Logit, and Probit Models*. Beverly Hills: Sage Publications, 1984.
- Allenby, Greg M. and Peter E. Rossi (1999) "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 57-78.
- Allison, P., "Problems with Fixed-Effects Negative Binomial Models," Manuscript, Department of Sociology, University of Pennsylvania, 2000.
- Allison, P. and R. Waterman, "Fixed-Effects Negative Binomial Regression Models," Manuscript, Department of Sociology, University of Pennsylvania, 2002.
- Amemiya, T. *Advanced Econometrics*. Cambridge, Harvard University Press, 1985.
- Andersen, E., "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 1970, pp. 283-301.
- Angrist, J., "Estimation of Limited Dependent Variable Models with Binary Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19, 1, 2001, pp. 1-14.
- Avery, R., L. Hansen, and J. Hotz. "Multiperiod Probit Models and Orthogonality Condition Estimation." *International Economic Review*, 24, 1983, pp. 21-35.
- Beck, N., D. Epstein and S. Jackman, "Estimating Dynamic Time Series Cross Section Models with a Binary Dependent Variable," Manuscript, Department of Political Science, University of California, San Diego, 2001.
- Ben-Akiva, M., and S. Lerman. *Discrete Choice Analysis*. London: MIT Press, 1985.
- Berndt, E., B. Hall, R. Hall, and J. Hausman. "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement*, 3/4, 1974, pp. 653-665.
- Berry, S., J. Levinsohn and A. Pakes., "Automobile Prices in Market Equilibrium." *Econometrica*, 63, 4, 1995, pp. 841-890.
- Bertschek, I. and M. Lechner, "Convenient Estimators for the Panel Probit Model," *Journal of Econometrics*, 87, 2, 1998, pp. 329-372
- Bhat, C. "A Heteroscedastic Extreme Value Model of Intercity Mode Choice," *Transportation Research*, 30, 1995, 1, pp. 16-29.
- Bhat, C. "Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Mode Choice Modeling." Department of Civil Engineering, University of Massachusetts, Amherst, 1996.
- Bhat, C., "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," Manuscript, Department of Civil Engineering, University of Texas, Austin, 1999.
- Brant, R., "Assessing Proportionality in the Proportional Odds Model for Ordered Logistic Regression," *Biometrics*, 46, 1990, pp. 1171-1178.
- Breusch, T., and A. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, 47, 1979, pp. 1287-1294.
- Breusch, T., and A. Pagan. "The LM Test and Its Applications to Model Specification in Econometrics." *Review of Economic Studies*, 47, 1980, pp. 239-254.
- Boyes, W., D. Hoffman, and S. Low, "An Econometric Analysis of the Bank Credit Scoring Problem," *Journal of Econometrics*, 40, 1989, pp. 3-14.
- Butler, J. and P. Chatterjee, "Pet Econometrics: Ownership of Cats and Dogs," Working Paper 95-WP1, Department of Economics, Vanderbilt University, 1995.
- Butler, J. and P. Chatterjee, "Tests of the Specification of Univariate and Bivariate Ordered Probit," *Review of Economics and Statistics*, 79, 1997, pp. 343-347.

- Butler, J., Finegan, T. and J. Siegfried, "Does More Calculus Improve Student Learning in Intermediate Micro- and Macroeconomic Theory?" *Journal of Applied Econometrics*, 13, 2, 1998m pp. 185-202.
- Butler, J., and R. Moffitt. "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model." *Econometrica*, 50, 1982, pp. 761-764.
- Calhoun, C., "Desired and Excess Fertility in Europe and the United States: Indirect Estimates from World Fertility Survey Data," *European Journal of Population*, 7, 1991, pp. 29-57.
- Cameron, A., and P. Trivedi. "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests." *Journal of Applied Econometrics*, 1, 1986, pp. 29-54.
- Cameron, C., and P. Trivedi. *Regression Analysis of Count Data*. New York: Cambridge University Press, 1998.
- Cameron, C. and F. Windmeijer, "R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization," Working Paper No. 93-24, Department of Economics, University of California, Davis, 1993.
- Caudill, S. "An Advantage of the Linear Probability Model Over Probit or Logit." *Oxford Bulletin of Economics and Statistics*, 50, 1988, pp. 425-427.
- Cecchetti, S., "The Frequency of Price Adjustment: A Study of the Newsstand Prices of Magazines," *Journal of Econometrics*, 31, 3, 1986, pp. 255-274.
- Chamberlain, G. "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 1980, pp. 225-238.
- Charlier, E., B. Melenberg and A. Van Soest, "A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model with an Application to Labor Force Participation," *Statistica Neerlandica*, 49, 1995, pp. 324-343.
- Chesher, A., and M. Irish. "Residual Analysis in the Grouped Data and Censored Normal Linear Model." *Journal of Econometrics*, 34, 1987, pp. 33-62.
- Christofides, L., T. Stengos and R. Swidinsky, "On the Calculation of Marginal Effects in the Bivariate Probit Model," *Economics Letters*, 54, 3, 1997, pp. 203-208.
- Christofides, L., T. Hardin, and R. Stengos, "On the Calculation of Marginal Effects in the Bivariate Probit Model: Corrigendum" *Economics Letters*, 68, 2000 pp. 339-340.
- Contoyannis, C., A. Jones and N. Rice, "The Dynamics of Health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19, 4, 2004, pp. 473-503.
- Cramer, J. "Predictive Performance of the Binary Logit Model in Unbalanced Samples." *Journal of the Royal Statistical Society, Series D (The Statistician)* 48, 1999, pp. 85-94.
- D'Addio, Eriksson, T. and P. Frijters, "An Analysis of the Determinants of Job Satisfaction when Individuals' Baseline Satisfaction Levels May Differ," Working Paper 2003-16, Center for Applied Microeconometrics, University of Copenhagen, 2003.
- Daganzo, C., *The Multinomial Probit Model: The Theory and Its Application to Demand Forecasting*, New York, Academic Press, 1979.
- Das M. and A. van Soest, "A Panel Data Model for Subjective Information on Household Income Growth." *Journal of Economic Behavior and Organization* 40, 2000, 409-426
- Dempster, A., N. Laird, and D. Rubin. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B*, 39, 1977, pp. 1-38.
- Econometric Software, Inc. *LIMDEP*, Plainview, NY, Version 9.0, 2007.
- Efron, B. "Regression and ANOVA with Zero-One Data: Measures of Residual Variation." *Journal of the American Statistical Association*, 73, 1978, pp. 113-212.
- Eisenberg, D. and B. Rowe, "The Effect of Serving in the Vietnam War on Smoking Behavior Later in Life," manuscript, School of Public Health, University of Michigan, 2006.
- Fabbri, D., C. Monfardini and R. Radice, "Testing Endogeneity in the Bivariate Probit Model: Monte Carlo Evidence and an Application to Health Economics," Manuscript, Department of Economics, University of Bologna, 2004.
- Ferrer-i-Carbonel A. and P. Frijter,, "The Effect of Methodology on the Determinants of Happiness", *Economic Journal*, 114, 2004, pp. 715-719.
- Fernandez, A., and J. Rodriguez-Poo. "Estimation and Testing in Female Labor Participation Models: Parametric and Semiparametric Models," *Econometric Reviews*, 16, 1997, pp. 229-248.
- Freedman, D., "On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors"," *The American Statistician*, 60, 4, 2006, pp. 299-302.

- Frijters P., J. Haisken-DeNew and M. Shields, The Value of Reunification in Germany: An Analysis of Changes in Life Satisfaction, *Journal of Human Resources*, 39, 3, 2004, pp. 649-674.
- Gandelman, N., "Homeownership and Gender," Manuscript, Universidad ORT, Uruguay, 2005.
- Gerfin, M. "Parametric and Semi-Parametric Estimation of the Binary Response Model." *Journal of Applied Econometrics*, 11, 1996, pp. 321-340.
- Geweke, J., M. Keane, and D. Runkle. "Alternative Computational Approaches to Inference in the Multinomial Probit Model." *Review of Economics and Statistics*, 76, 1994, pp. 609-632.
- Geweke, J., Contemporary Bayesian Econometrics and Statistics, John Wiley and Sons, New York, 2005.
- Gaudry, M. and M. Dagenais, "The Dogit Model," *Transportation Research, Series B*, 13, 1979, pp. 105-111.
- Goldberger, A., *Functional Form and Utility: A Review of Consumer Demand Theory*. Westview Press, Boulder, Colorado, 1987.
- Gourieroux, C., and A. Monfort. *Simulation-Based Methods Econometric Methods*. Oxford: Oxford University Press, 1996.
- Greene, W., "A Statistical Model for Credit Scoring," Stern School of Business, Dept. of Economics, Working Paper 92-29, 1992.
- Greene, W. "Sample Selection in the Poisson Regression Model." Working Paper No. EC-95-6, Department of Economics, Stern School of Business, New York University, 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. "FIML Estimation of Sample Selection Models for Count Data." Working Paper No. 97-02, Department of Economics, Stern School of Business, New York University, 1997.
- Greene, W. "Gender Economics Courses in Liberal Arts Colleges: Further Results." *Journal of Economic Education*, 29, 4, 1998, pp. 291-300.
- Greene, W., "Fixed and Random Effects in Nonlinear Models," Working Paper EC-01-01, Stern School of Business, Department of Economics, 2001.
- Greene, W., "Convenient Estimators for the Panel Probit Model" *Empirical Economics*, 29, 1, 2004a, pp. 21-47.
- Greene, W. "Fixed Effects and the Incidental Parameters Problem in the Tobit Model," *Econometric Reviews*, 23, 2, 2004b, pp. 125-148.
- Greene, W., "Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model," *Econometric Reviews*, 2004b, vol. 23, issue 2, pages 125-147
- Greene, W., "Censored Data and Truncated Distributions," in T. Mills and K. Patterson, eds., Palgrave Handbook of Econometrics, Volume 1: Econometric Theory, Palgrave, Hampshire, 2006a.
- Greene, W., "A Method of Incorporating Sample Selection in a Nonlinear Model," Working Paper 06-10, Stern School of Business, New York University, 2006b.
- Greene, W., *LIMDEP, User's Manual*, Econometric Software, Plainview, NY, 2007a.
- Greene W., "Functional Form and Heterogeneity in Models for Count Data, Department of Economics, Stern School of Business, New York University, Working Paper 07-10, 2007b.
- Greene, W., *Econometric Analysis*, 6<sup>th</sup> ed., Prentice Hall, Upper Saddle River, 2008.
- Greene, W., M. Harris, B. Hollingsworth, and P. Maitra, "A Bivariate Latent Class Correlated Generalized Ordered Probit Model with an Application to Modeling Observed Obesity Levels," Manuscript, Department of Econometrics and Business Statistics, Monash University, Melbourne, 2007.
- Greene and D. Hensher, "Accounting for Heterogeneity in the Variance of Unobserved Effects in Mixed Logit Models," *Transportation Research, B: Methodology*, 40, 1, 2006, pp. 75-92.
- Greene, W., S. Rhine and M. Toussaint-Comeau, "The Importance of Check-Cashing Businesses to the Unbanked: A Look at Racial /Ethnic Differences," *Review of Economics and Statistics*, 88, 1, 2006, pp. 146-157
- Groot and van den Brink, 2003
- Gurmu, S., "Semi-Parametric Estimation of Hurdle Regression Models with an Application to Medicaid Utilization," *Journal of Applied Econometrics*, 12, 3, 1997, pp. 225-242.
- Hardle, W., and C. Manski, ed., "Nonparametric and Semiparametric Approaches to Discrete Response Analysis." *Journal of Econometrics*, 58, 1993, pp. 1-274.
- Harris, M. and X. Zhao, "Modelling Tobacco Consumption with a Zero Inflated Ordered Probit Model," Working Paper 14/04, Monash University School of Business and Economics, 2007.

- Harvey, A. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica*, 44, 1976, pp. 461–465.
- Hausman, J., B. Hall, and Z. Griliches. "Economic Models for Count Data with an Application to the Patents–R&D Relationship." *Econometrica*, 52, 1984, pp. 909–938.
- Heckman, J. "Sample Selection Bias as a Specification Error." *Econometrica*, 47, 1979, pp. 153–161.
- State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'INSEE*, 30, 1978, pp. 227–269.
- Heckman, J. "Sample Selection Bias as a Specification Error." *Econometrica*, 47, 1979, pp. 153–161.
- Heckman, J., "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, 1981a.
- Heckman, J., "Heterogeneity and State Dependence," in *Studies of Labor Markets*, edited by S. Rosen, NBER, University of Chicago Press, Chicago, 1981b.
- Heckman, J., "Heterogeneity and State Dependence," in S. Rosen, ed. *Studies in Labor Markets*, University of Chicago Press, Chicago, 1981c.
- Heckman, J. and T. MaCurdy, T. "A Life Cycle Model of Female Labor Supply" *Review of Economic Studies* 47, 1981, pp. 247–283.
- Heckman, J. and J. Snyder, "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators," *Rand Journal of Economics*, 28, 0, 1997.
- Hensher, D. and W. Greene, "Specification and Estimation of the Nested Logit Model: Alternative Normalizations," *Transportation Research B*, 1999, July.
- Hensher, D. and Greene, W., "The Mixed Logit Model: The State of Practice" with David Hensher, *Transportation Research, B*, 30, 2003, pp. 133–176.
- Hensher, D. and W. Greene, "Accounting for Heterogeneity in the Variance of Unobserved Effects in Mixed Logit Models," *Transportation Research, B: Methodology*, 40, 1, 2006, pp. 75–92.
- Hensher, D., J. Rose. and W. Greene, *Applied Choice Analysis*, Cambridge University Press, 2005.
- Honore, B. and E. Kyriazidou, "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 4, 2000a, pp. 839–874.
- Honore, B. and E. Kyriazidou, "Estimation of Tobit-Type Models with Individual Specific Effects," *Econometric Reviews*, 19, 3, 2000b, pp. 341–366.
- Honore, B., "Non-Linear Models with Panel Data," Institute For Fiscal Studies, CEMMAP, Working Paper CWP13/02, 2002.
- Horowitz, J., "A Smoothed Maximum Score Estimator for the Binary Response Model, *Econometrica*, 60, 1992, pp. 505–531.
- Horowitz, J. "Semiparametric Estimation of a Work-Trip Mode Choice Model." *Journal of Econometrics*, 58, 1993, pp. 49–70.
- Hunt, G., "Alternative Nested Logit Model Structures and the Special Case of Partial Degeneracy," *Journal of Regional Science*, 40, February, 2000, pp. 89–113.
- Hsiao, C., *Analysis of Panel Data*, Cambridge University Press, Cambridge, 1986.
- Hsiao, C., *Analysis of Panel Data, 2<sup>nd</sup> Ed.* Cambridge University Press, Cambridge, 2003.
- Hujer, R. and H. Schneider, "The Analysis of Labor Market Mobility Using Panel Data," *European Economic Review*, 33, 1989, pp. 530–536.
- Hyslop, D., "State Dependence, Serial Correlation, and Heterogeneity in Labor Force Participation of Married Women," *Econometrica*, 67, 6, 1999, pp. 1255–1294.
- Jain, D., N. Vilcassim, and P. Chintagunta. "A Random-Coefficients Logit Brand Choice Model Applied to Panel Data." *Journal of Business and Economic Statistics*, 12, 3, 1994, pp. 317–328.
- Jakubson, G., "The Sensitivity of Labor Supply Parameters to Unobserved Individual Effects: Fixed and Random Effects Estimates in a Nonlinear Model Using Panel Data," *Journal of Labor Economics*, 6, 1988, pp. 302–329
- Jones, J. and J. Landwehr, "Removing Heterogeneity Bias from Logit Model Estimation," *Marketing Science*, 7,1, 1988, pp. 41–59.
- Kassouf, A. and R. Hoffmann, "Work Related Injuries Involving Children and Adolescents: Application of a Recursive Bivariate Probit Model," *Brazilian Review of Econometrics*, 26, 1, 2006, pp. 105–126.
- Katz, E., "Bias in Conditional and Unconditional Fixed Effects Logit Estimation," *Political Analysis*, 9, 4, 2001, pp. 379–384.
- Kay, R., and S. Little. "Assessing the Fit of the Logistic Model: A Case Study of Children with Haemolytic Uraemic Syndrome." *Applied Statistics*, 35, 1986, pp. 16–30.

- Keane, M., "Simulation Estimators for Panel Data Models with Limited Dependent Variables," in Maddala, G. and C. Rao, eds., *Handbook of Statistics*, Volume 11, Chapter 20, North Holland, Amsterdam, 1993.
- Kiefer, N. "Testing for Independence in Multivariate Probit Models." *Biometrika*, 69, 1982, pp. 161–166.
- King, G., "A Seemingly Unrelated Poisson Regression Model," *Sociological Methods and Research*, 17, 3, 1989, pp. 235-255.
- Klein, R. and R. Spady, "An Efficient Semiparametric Estimator for Discrete Choice," *Econometrica*, 61, 1993, pp. 387-421.
- Koop, G., *Bayesian Econometrics*, John Wiley and Sons, New York, 2003
- Krinsky, I. and L. Robb. "On Approximating the Statistical Properties of Elasticities." *Review of Economics and Statistics*, 68, 4, 1986, pp. 715-719.
- Kyriazidou, E., "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 6, 1997, pp.; 1335-1364.
- Kyriazidou, E., "Estimation of Dynamic Panel Data Sample Selection Models," *Review of Economic Studies*, 68, 2001, pp. 543-572.
- Lambert, D., "Zero Inflated Poisson Regression, with an Application to Defects in Manufacturing," *Technometrics*, 34, 1, 1992, pp. 1-14.
- Lancaster, T. 2000. The incidental parameters problem since 1948. *Journal of Econometrics*, **95**: 391-414.
- Lancaster, T., *An Introduction to Modern Bayesian Inference*, Oxford University Press, Oxford, 2004
- Lee, M., "A Root-N Consistent Semiparametric Estimator for Related-Effects Binary Response Panel Data," *Econometrica*, 1967, pp. 427-434.
- Lee, E., J. Lee and D. Eastwood, "A Two Step Estimation of Consumer Adoption of Technology Based Service Innovations," *Journal of Consumer Affairs*, 37, 2, 2003, pp. 37-62.
- Lerman, S. and Manski, C., "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in Manski, C. and McFadden, D. (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, Massachusetts, 1981.
- Lewbel, A., Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 1, 2000, pp. 145-177.
- Li, Q and J. Racine, *Nonparametric Econometrics*, Princeton University Press, Princeton, 2007.
- Long, S. *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks, CA, 1997.
- Maddala, G. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York, 1983.
- Magee, L., J. Burbidge and L. Robb, 2000. "The Correlation Between Husband's and Wife's Education: Canada. 1971-1996" *Social and Economic Dimensions of an Aging Population Research Papers*, 24, McMaster University, 2000.
- Magnac, T., "State Dependence and Heterogeneity in Youth Unemployment Histories," *Working Paper, INRA and CREST*, Paris, 1997.
- Manski, C. "The Maximum Score Estimator of the Stochastic Utility Model of Choice." *Journal of Econometrics*, 3, 1975, pp. 205–228.
- Manski, C. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics*, 27, 1985, pp. 313–333.
- Manski, C. "Operational Characteristics of the Maximum Score Estimator." *Journal of Econometrics*, 32, 1986, pp. 85–100.
- Manski, C., "Semiparametric Analysis of the Random Effects Linear Model from Binary Response Data," *Econometrica*, 55, 1987, pp. 357-362.
- Manski, C., and S. Lerman. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica*, 45, 1977, pp. 1977–1988.
- Manski, C., and S. Thompson. "MSCORE: A Program for Maximum Score Estimation of Linear Quantile Regressions from Binary Response Data." Mimeo, University of Wisconsin, Madison, Department of Economics, 1986.
- Matzkin, R. "Nonparametric Identification and Estimation of Polytomous Choice Models." *Journal of Econometrics*, 58, 1993, pp. 137–168.
- McFadden, D. "Econometric Models of Probabilistic Choice," in Manski, C. and McFadden, (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, Mass., 1981

- McFadden, D., "Economic Choices," *American Economic Review*, 93, 3, 2001, pp. 351-378.
- McFadden, D. and K. Train, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 2000, pp. 447-470.
- McQuestion, M., "A Bivariate Probit Analysis of Social Interaction and Treatment Effects," Center for Demography and Ecology, University of Wisconsin, Working Paper 2000-05, 2000.
- Mullahy, J. "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics*, 33, 1986, pp. 341-365.
- Mundlak, Y. "On the Pooling of Time Series and Cross Sectional Data." *Econometrica*, 56, 1978, pp. 69-86.
- Munkin, M. and P. Trivedi, "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application," *Econometrics Journal*, 2, 1999, pp. 29-49.
- Murphy, K., and R. Topel. "Estimation and Inference in Two Step Econometric Models." *Journal of Business and Economic Statistics*, 3, 1985, pp. 370-379.
- Newey, W., "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 36, 1987, pp. 231-250.
- Newey, W., J. Powell, and J. Walker. "Semiparametric Estimation of Selection Models." *American Economic Review*, 80, 1990, pp. 324-328.
- Neyman, J. and E. Scott., "Consistent Estimates Based on Partially Consistent Observations," *Econometrica* 16: 1948.1-32.
- Olsen, R., "A Note on the Uniqueness of the Maximum Likelihood Estimator of the Tobit Model," *Econometrica*, 46, 1978, pp. 1211-1215.
- Pudney, S. and M. Shields, "Gender, Race, Pay and Promotion in the British Nursing Profession: Estimation of a Generalized Ordered Probit Model," *Journal of Applied Econometrics*, 15, 4, 2000, pp. 367-399.
- Pratt, J., "Concavity of the Log Likelihood," *Journal of the American Statistical Association*, 76, 1981, pp. 103-106.
- Rabe-Hesketh, S., Skrondal, A. and Pickles, A.. "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics*, 128, 2, 2005, pp. 301-323.
- Rasch, G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmark Paedogiska, Copenhagen, 1960.
- Revelt, D. and K. Train, "Mixed Logit with Repeated Choices: Households' Choice of Appliance Efficiency Level," *Review of Economics and Statistics*, 80, 4, 1998, pp. 647-657.
- Riphahn, R., A. Wambach, and A. Million, A., "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" *Journal of Applied Econometrics*, Vol. 18, No. 4, 2003, pp. 387-405
- Rossi, P. and G. Allenby, "Bayesian Statistics and Marketing," *Marketing Science*, 22, 2003, 304-328.
- Samuelson, P. *Foundations of Economic Analysis*, Atheneum Press, 1947.
- Sepanski, J., "On a Random Coefficients Probit Model," *Communications in Statistics – Theory and Methods*, 29, 2000, pp. 2493-2505.
- Shaw, D. "On-Site Samples' Regression Problems of Nonnegative Integers, Truncation, and Endogenous Stratification." *Journal of Econometrics*, 37, 1988, pp. 211-223.
- Silva, J., "A Score Test for Non-Nested Hypotheses with Applications to Discrete Response Models," *Journal of Applied Econometrics*, 16, 5, 2001, pp. 577-598.
- Stata. *Stata User's Guide, Version 9*. College Station, TX: Stata Press 2006.
- Terza, J. "Ordinal Probit: A Generalization." *Communications in Statistics*, 14, 1985a, pp. 1-12.
- Terza, J. "A Tobit Type Estimator for the Censored Poisson Regression Model." *Economics Letters*, 18, 1985b, pp. 361-365.
- Terza, J. "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects." *Journal of Econometrics*, 84, 1, 1998, pp. 129-154.
- Tobias, J. and M. Li, "Calculus Attainment and Grades Received in Intermediate Economic Theory," *Journal of Applied Econometrics*, 21, 6, 2006, pp. 893-896.
- Train, K. "Halton Sequences for Mixed Logit," Manuscript, Department of Economics, University of California, Berkeley, 1999.
- Train, K., *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, 2003.

- E Van Doorslaer and Nonneman, W, "Economic incentives in the health care industry: implications for health policy making," *Health Policy*, 7, 2, 1987, pp. 109-114
- Vella, F. and M. Verbeek, "Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias," *Journal of Econometrics*, 90, 1999, pp. 239-263.
- Vuong, Q. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." *Econometrica*, 57, 1989, pp. 307-334.
- Wagstaff, A., "The Demand for Health: An Empirical Reformulation of the Grossman Model," *Health Economics*, 2, 1993, pp. 189-198.
- White, H. "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity." *Econometrica*, 48, 1980, pp. 817-838.
- White, N. and A. Wolaver, "Occupation Choice, Information and Migration," *Review of Regional Studies*, 33, 2, 2003, pp. 142-163.
- Willis, J., "Magazine Prices Revisited," *Journal of Applied Econometrics*, 21,3, 2006, pp. 337-344.
- Winkelmann, R., *Econometric Analysis of Count Data*, Springer Verlag, Heidelberg, 4<sup>th</sup> ed. 2003.
- Winkelmann, R., "Health Care Reform and the Number of Doctor Visits – An Econometric Analysis", *Journal of Applied Econometrics*, Vol. 19, No. 4, 2004, pp. 455-472.
- Wooldridge, J., "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68, 1, 1995, pp. 115-132.
- Wooldridge, J., *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, 2002.
- Wooldridge, J., "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, 1, 2005, pp. 39-54.
- Wooldridge, J., "Simple Solutions to the Initial Conditions Problem in Dynamic Nonlinear Panel Data Models with Unobserved Heterogeneity," CEMMAP Working Paper CWP18/02, Centre for Microdata and Practice, IFS and University College, London, 2002.
- Wynand, P. and B. van Praag, "The Demand for Deductibles in Private Health Insurance," *Journal of Econometrics*, 17, 1981, pp. 229-252.
- Zavoina, R. and W. McElvey, "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology*, Summer, 1975, pp. 103-120.