CHAPTER 7

Discrete Data

7.1 INTRODUCTION

In this chapter, we consider situations in which an analyst has at his disposal a random sample of N individuals, having recorded histories indicating the presence or absence of an event in each of T equally spaced discrete time periods. Statistical models in which the endogenous random variables take only discrete values are known as discrete, categorical, qualitative-choice, or quantal-response models. The literature, both applied and theoretical, on this subject is vast. Amemiya (1981), Maddala (1983), and McFadden (1976, 1984) have provided excellent surveys. Thus, the focus of this chapter will be only on controlling for unobserved characteristics of individual units to avoid specification bias. Many important and more advanced topics are omitted, such as continuous-time and duration-dependence models (Chamberlain (1978b); Flinn and Heckman (1982); Heckman and Borjas (1980); Heckman and Singer (1982); Lancaster (1990); Nickell (1979); Singer and Spilerman (1976)).

7.2 SOME DISCRETE-RESPONSE MODELS

In this section, we briefly review some widely used discrete-response models. We first consider the case in which the dependent variable y can assume only two values, which for convenience and without any loss of generality will be the value 1 if an event occurs and 0 if it does not. Examples of this include purchases of durables in a given year, participation in the labor force, the decision to enter college, and the decision to marry.

The discrete outcome of y can be viewed as the observed counterpart of a latent continuous random variable crossing a threshold. Suppose that the continuous latent random variable, y^*, is a linear function of a vector of explanatory variables, $\mathbf{x}$,

$$y^* = \boldsymbol{\beta}'\mathbf{x} + v, \tag{7.2.1}$$

where the error term v is independent of $\mathbf{x}$ with mean zero. Suppose, instead of

observing y^*, we observe y, where

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{if } y^* \leq 0. \end{cases} \tag{7.2.2}$$

Then the expected value of y_i is the probability that the event will occur,

$$\begin{aligned} E(y \mid \mathbf{x}) &= 1 \cdot \Pr(v > -\boldsymbol{\beta}'\mathbf{x}) + 0 \cdot \Pr(v \leq -\boldsymbol{\beta}'\mathbf{x}) \\ &= \Pr(v > -\boldsymbol{\beta}'\mathbf{x}) \\ &= \Pr(y = 1 \mid \mathbf{x}). \end{aligned} \tag{7.2.3}$$

When the probability law for generating v follows a two-point distribution $(1 - \boldsymbol{\beta}'\mathbf{x})$ and $(-\boldsymbol{\beta}'\mathbf{x})$, with probabilities $\boldsymbol{\beta}'\mathbf{x}$ and $(1 - \boldsymbol{\beta}'\mathbf{x})$, respectively, we have the linear-probability model

$$y = \boldsymbol{\beta}'\mathbf{x} + v, \tag{7.2.4}$$

with $Ev = \boldsymbol{\beta}'\mathbf{x}(1 - \boldsymbol{\beta}'\mathbf{x}) + (1 - \boldsymbol{\beta}'\mathbf{x})(-\boldsymbol{\beta}'\mathbf{x}) = 0$. When the probability density function of v is a standard normal density function, $(1/\sqrt{2\pi}) \times \exp(-v^2/2) = \phi(v)$, we have the probit model,

$$\begin{aligned} \Pr(y = 1 \mid \mathbf{x}) &= \int_{-\boldsymbol{\beta}'\mathbf{x}}^{\infty} \phi(v)\, dv \\ &= \int_{-\infty}^{\boldsymbol{\beta}'\mathbf{x}} \phi(v)\, dv = \Phi(\boldsymbol{\beta}'\mathbf{x}). \end{aligned} \tag{7.2.5}$$

When the probability density function is a standard logistic,

$$\frac{\exp(v)}{(1 + \exp(v))^2} = [(1 + \exp(v))(1 + \exp(-v))]^{-1},$$

we have the logit model

$$\Pr(y = 1 \mid \mathbf{x}) = \int_{-\boldsymbol{\beta}'\mathbf{x}}^{\infty} \frac{\exp(v)}{(1 + \exp(v))^2}\, dv_i = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}. \tag{7.2.6}$$

Let $F(\boldsymbol{\beta}'\mathbf{x}) = E(y_i \mid \mathbf{x})$. Then the three commonly used parametric models for the binary choice may be summarized with a single index w as follows:

Linear-probability model,

$$F(w) = w. \tag{7.2.7}$$

Probit model,

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du = \Phi(w). \tag{7.2.8}$$

Logit model,

$$F(w) = \frac{e^w}{1 + e^w}. \tag{7.2.9}$$

The linear-probability model is a special case of the linear regression model with heteroscadastic variance, $\boldsymbol{\beta}'\mathbf{x}(1-\boldsymbol{\beta}'\mathbf{x})$. It can be estimated by least squares or weighted least squares (Goldberger (1964)). But it has an obvious defect in that $\boldsymbol{\beta}'\mathbf{x}$ is not constrained to lie between 0 and 1 as a probability should, whereas in the probit and logit models it is.

The probability functions used for the probit and logit models are the standard normal distribution and the logistic distribution, respectively. We use cumulative standard normal because in the dichotomy case there is no way to identify the variance of a normal density. The logit probability density function is symmetric around 0 and has a variance of $\pi^2/3$. Because they are distribution functions, the probit and logit models are bounded between 0 and 1.

The cumulative normal distribution and the logistic distribution are very close to each other; the logistic distribution has sightly heavier tails (Cox (1970)). Moreover, the cumulative normal distribution Φ is reasonably well approximated by a linear function for the range of probabilities between 0.3 and 0.7. Amemiya (1981) has suggested an approximate conversion rule for the coefficients of these models. Let the coefficients for the linear-probability, probit, and logit models be denoted as $\hat{\boldsymbol{\beta}}_{\mathrm{LP}}$, $\hat{\boldsymbol{\beta}}_{\Phi}$, $\hat{\boldsymbol{\beta}}_{L}$, respectively. Then

$$\begin{aligned}&\hat{\boldsymbol{\beta}}_{L} \simeq 1.6\hat{\boldsymbol{\beta}}_{\Phi},\\ &\hat{\boldsymbol{\beta}}_{\mathrm{LP}} \simeq 0.4\hat{\boldsymbol{\beta}}_{\Phi} \text{ except for the constant term,}\end{aligned} \tag{7.2.10}$$

and

$$\hat{\boldsymbol{\beta}}_{\mathrm{LP}} \simeq 0.4\hat{\boldsymbol{\beta}}_{\Phi} + 0.5 \text{ for the constant term.}$$

For a random sample of N individuals, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, the likelihood function for these three models can be written in general form as

$$L = \prod_{i=1}^{N} F(\boldsymbol{\beta}'\mathbf{x}_i)^{y_i}[1 - F(\boldsymbol{\beta}'\mathbf{x}_i)]^{1-y_i}. \tag{7.2.11}$$

Differentiating the logarithm of the likelihood function yields the vector of first derivatives and the matrix of second-order derivatives as

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{y_i - F(\boldsymbol{\beta}'\mathbf{x}_i)}{F(\boldsymbol{\beta}'\mathbf{x}_i)[1 - F(\boldsymbol{\beta}'\mathbf{x}_i)]} F'(\boldsymbol{\beta}'\mathbf{x}_i)\mathbf{x}_i, \tag{7.2.12}$$

and

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \Bigg\{ &-\sum_{i=1}^{N}\left[\frac{y_i}{F^2(\boldsymbol{\beta}'\mathbf{x}_i)} + \frac{1-y_i}{[1-F(\boldsymbol{\beta}'\mathbf{x}_i)]^2}\right][F'(\boldsymbol{\beta}'\mathbf{x}_i)]^2 \\ &+ \sum_{i=1}^{N}\left[\frac{y_i - F(\boldsymbol{\beta}'\mathbf{x}_i)}{F(\boldsymbol{\beta}'\mathbf{x}_i)[1-F(\boldsymbol{\beta}'\mathbf{x}_i)]}\right] F''(\boldsymbol{\beta}'\mathbf{x}_i)\Bigg\} \mathbf{x}_i\mathbf{x}_i',\end{aligned} \tag{7.2.13}$$

where $F'(\boldsymbol{\beta}'\mathbf{x}_i)$ and $F''(\boldsymbol{\beta}'\mathbf{x}_i)$ denote the first and second derivatives of $F(\boldsymbol{\beta}'\mathbf{x}_i)$ with respect to $\boldsymbol{\beta}'\mathbf{x}_i$. If the likelihood function (7.2.11) is concave, as in the

models discussed here (e.g., Amemiya (1985, p. 273)), then a Newton–Raphson method,

$$\hat{\boldsymbol{\beta}}^{(j)} = \hat{\boldsymbol{\beta}}^{(j-1)} - \left(\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right)^{-1}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(j-1)}} \left(\frac{\partial \log L}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(j-1)}}, \tag{7.2.14}$$

or a method of scoring,

$$\hat{\boldsymbol{\beta}}^{(j)} = \hat{\boldsymbol{\beta}}^{(j-1)} - \left[E\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right]^{-1}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(j-1)}} \left(\frac{\partial \log L}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(j-1)}}, \tag{7.2.15}$$

can be used to find the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ with arbitrary initial values $\hat{\boldsymbol{\beta}}^{(0)}$, where $\hat{\boldsymbol{\beta}}^{(j)}$ denotes the jth iterative solution.

In the case in which there are repeated observations of y for a specific value of $\mathbf{x}$, the proportion of $y = 1$ for individuals with the same characteristic $\mathbf{x}$ is a consistent estimator of $p = F(\boldsymbol{\beta}'\mathbf{x})$. Taking the inverse of this function yields $F^{-1}(p) = \boldsymbol{\beta}'\mathbf{x}$. Substituting $\hat{p}$ for p, we have $F^{-1}(\hat{p}) = \boldsymbol{\beta}'\mathbf{x} + \zeta$, where ζ denotes the approximation error of using $F^{-1}(\hat{p})$ for $F^{-1}(p)$. Since ζ has a nonscalar covariance matrix, we can apply the weighted least-squares method to estimate $\boldsymbol{\beta}$. The resulting estimator, which is generally referred to as the minimum-chi-square estimator, has the same asymptotic efficiency as the MLE and computationally may be simpler than the MLE. Moreover, in finite samples, the minimum-chi-square estimator may even have a smaller mean squared error than the MLE (e.g., Amemiya (1974, 1976, 1980b); Berkson (1944, 1955, 1957, 1980); Ferguson (1958); Neyman (1949)). However, despite its statistical attractiveness, the minimum-chi-square method is probably less useful than the maximum likelihood method in analyzing survey data than it is in the laboratory setting. Application of the minimum-chi-square method requires repeated observations for each value of the vector of explanatory variables. In survey data, most explanatory variables are continuous. The survey sample size has to be extremely large for the possible configurations of explanatory variables. Furthermore, if the proportion of $y = 1$ is 0 or 1 for a given $\mathbf{x}$, the minimum-chi-square method is not defined, but the maximum likelihood method can still be applied. For this reason, we shall confine our attention to the maximum likelihood method.[1]

When the dependent variable y_i can assume more than two values, things are more complicated. We can classify these cases into ordered and unordered variables. An example of ordered variables is

$$y_i = \begin{cases} 0 & \text{if the price of a home bought} < \$49{,}999, \\ 1 & \text{if the price of a home bought is } \$50{,}000\text{–}\$99{,}999, \\ 2 & \text{if the price of a home bought} > \$100{,}000. \end{cases}$$

An example of unordered variables is

$$y_i = \begin{cases} 1 & \text{if mode of transport is car,} \\ 2 & \text{if mode of transport is bus,} \\ 3 & \text{if mode of transport is train.} \end{cases}$$

In general, ordered models are used whenever the values taken by the discrete random variable y_i correspond to the intervals within which a continuous latent random variable y_i^* falls. Unordered models are used when more than one latent continuous random variable is needed to characterize the responses of y_i.

Assume that the dependent variable y_i takes $m_i + 1$ values $0, 1, 2, \ldots, m_i$ for the ith unit. To simplify the exposition without having to distinguish ordered from unordered models, we define $\sum_{i=1}^{N}(m_i + 1)$ binary variables as

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y_i = j, \quad i = 1, \ldots, N, \\ 0 & \text{if} \quad y_i \neq j, \quad j = 0, 1, \ldots, m_i. \end{cases} \tag{7.2.16}$$

Let $\text{Prob}(y_{ij} = 1 \mid \mathbf{x}_i) = F_{ij}$. We can write the likelihood function as

$$L = \prod_{i=1}^{N} \prod_{j=0}^{m_i} F_{ij}^{y_{ij}}. \tag{7.2.17}$$

The complication in the multivariate case is in the specification of F_{ij}. Once F_{ij} is specified, general results concerning the methods of estimation and their asymptotic distributions for the dichotomous case also apply here. However, contrary to the univariate case, the similarity between the probit and logit specifications no longer holds. In general, they will lead to different inferences.

The multivariate probit model follows from the assumption that the errors of the latent response functions across alternatives are multivariate normally distributed. Its advantage is that it allows the choice among alternatives to have arbitrary correlation. Its disadvantage is that the evaluation of $\text{Prob}(y_i = j)$ involves multiple integrations, which can be computationally infeasible.

The conditional logit model follows from the assumption that the errors of the latent response functions across alternatives are independently, identically distributed with type I extreme value distribution (McFadden (1974)). Its advantage is that the evaluation of $\text{Prob}(y_i = j)$ does not involve multiple integration. Its disadvantage is that the relative odds between two alternatives are independent of the presence or absence of the other alternatives – the so-called independence of irrelevant alternatives. If the errors among alternatives are not independently distributed, this can lead to grossly false predictions of the outcomes. For discussion of model specification tests, see Hausman and McFadden (1984), Hsiao (1992b), Lee (1982, 1987), and Small and Hsiao (1985).

Because in many cases a multiresponse model can be transformed into a dichotomous model characterized by the $\sum_{i=1}^{N}(m_i + 1)$ binary variables as in (7.2.16),[2] for ease of exposition we shall concentrate on the dichotomous model.[3]

When there is no information about the probability laws for generating v_i, a semiparametric approach can be used to estimate $\boldsymbol{\beta}$ subject to a certain normalization rule (e.g., Klein and Spady (1993); Manski (1985); Powell, Stock, and Stoker (1989)). However, whether an investigator takes a parametric or semiparametric approach, the cross-sectional model assumes that the error term v_i in the latent response function (7.2.1) is independently, identically distributed and is independent of $\mathbf{x}_i$. In other words, conditional on $\mathbf{x}_i$, everyone has the same

probability that an event will occur. It does not allow the possibility that the average behavior given $\mathbf{x}$ can be different from individual probabilities, that is, that it does not allow $\Pr(y_i = 1 \mid \mathbf{x}) \neq \Pr(y_j = 1 \mid \mathbf{x})$. The availability of panel data provides the possibility of distinguishing average behavior from individual behavior by decomposing the error term v_{it} into

$$v_{it} = \alpha_i + \lambda_t + u_{it}, \tag{7.2.18}$$

where α_i and λ_t denote the effects of omitted individual-specific and time-specific variables, respectively. In this chapter we shall demonstrate the misspecifications that can arise because of failure to control for unobserved characteristics of the individuals in panel data, and discuss possible remedies.

7.3 PARAMETRIC APPROACH TO STATIC MODELS WITH HETEROGENEITY

Statistical models developed for analyzing cross-sectional data essentially ignore individual differences and treat the aggregate of the individual effect and the omitted-variable effect as a pure chance event. However, as stated in Chapter 1, a discovery of a group of married women having an average yearly labor participation rate of 50 percent could lead to diametrically opposite inferences. At one extreme, each woman in a homogeneous population could have a 50 percent chance of being in the labor force in any given year, whereas at the other extreme 50 percent of women in a heterogeneous population might always work and 50 percent never work. Either explanation is consistent with the given cross-sectional data. To discriminate among the many possible explanations, we need information on individual labor-force histories in different subintervals of the life cycle. Panel data, through their information on intertemporal dynamics of individual entities, provide the possibility of separating a model of individual behavior from a model of the average behavior of a group of individuals.

For simplicity, we shall assume that the heterogeneity across cross-sectional units is time-invariant,[4] and these individual-specific effects are captured by decomposing the error term v_{it} in (7.2.1) as $\alpha_i + u_{it}$. When the α_i are treated as fixed, $\operatorname{Var}(v_{it} \mid \alpha_i) = \operatorname{Var}(u_{it}) = \sigma_u^2$. When they are treated as random, we assume that $E\alpha_i = Eu_{it} = E\alpha_i u_{it} = 0$ and $\operatorname{Var}(v_{it}) = \sigma_u^2 + \sigma_\alpha^2$. However, as discussed earlier, when the dependent variables are binary, the scale factor is not identifiable. Thus, for ease of exposition, we normalize the variance σ_u^2 of u to be equal to 1 for the specifications discussed in the rest of this chapter.

The existence of such unobserved permanent components allows individuals who are homogeneous in their observed characteristics to be heterogeneous in their response probabilities $F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$. For example, heterogeneity will imply that the sequential-participation behavior of a woman, $F(\boldsymbol{\beta}'\mathbf{x} + \alpha_i)$, within a group of observationally identical women differs systematically from $F(\boldsymbol{\beta}'\mathbf{x})$ or the average behavior of the group, $\int F(\boldsymbol{\beta}'\mathbf{x} + \alpha)\, dH(\alpha \mid \mathbf{x})$, where $H(\alpha \mid \mathbf{x})$ gives the population probability (or empirical distribution) for α conditional on $\mathbf{x}$.[5] In this section, we discuss statistical inference of the common parameters $\boldsymbol{\beta}$ based on a parametric specification of $F(\cdot)$.

7.3.1 Fixed-Effects Models

7.3.1.a Maximum Likelihood Estimator

If the individual-specific effect, α_i, is assumed to be fixed,[6] then both α_i and $\boldsymbol{\beta}$ are unknown parameters to be estimated for the model $\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i) = F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)$. When T tends to infinity, the MLE is consistent. However, T is usually small for panel data. There are only a limited number of observations to estimate α_i. Thus, we have the familiar incidental-parameter problem (Neyman and Scott (1948)). Any estimation of the α_i is meaningless if we intend to judge the estimators by their large-sample properties. We shall therefore concentrate on estimation of the common parameters, $\boldsymbol{\beta}$.

Unfortunately, contrary to the linear-regression case where the individual effects α_i can be eliminated by taking a linear transformation such as the first difference, in general no simple transformation exists to eliminate the incidental parameters from a nonlinear model. The MLEs for α_i and $\boldsymbol{\beta}$ are not independent of each other for the discrete-choice models. When T is fixed, the inconsistency of $\hat{\alpha}_i$ is transmitted into the MLE for $\boldsymbol{\beta}$. Hence, even if N tends to infinity, the MLE of $\boldsymbol{\beta}$ remains inconsistent.

We demonstrate the inconsistency of the MLE for $\boldsymbol{\beta}$ by considering a logit model. The log likelihood function for this model is

$$\log L = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log[1 + \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)] + \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i). \tag{7.3.1}$$

For ease of illustration, we consider the special case of $T = 2$ and one explanatory variable, with $x_{i1} = 0$ and $x_{i2} = 1$. Then the first-derivative equations are

$$\begin{aligned}\frac{\partial \log L}{\partial \beta} &= \sum_{i=1}^{N}\sum_{t=1}^{2}\left[-\frac{e^{\beta x_{it}+\alpha_i}}{1 + e^{\beta x_{it}+\alpha_i}} + y_{it}\right] x_{it} \\ &= \sum_{i=1}^{N}\left[-\frac{e^{\beta+\alpha_i}}{1 + e^{\beta+\alpha_i}} + y_{i2}\right] = 0, \end{aligned} \tag{7.3.2}$$

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{2}\left[-\frac{e^{\beta x_{it}+\alpha_i}}{1 + e^{\beta x_{it}+\alpha_i}} + y_{it}\right] = 0. \tag{7.3.3}$$

Solving (7.3.3), we have

$$\hat{\alpha}_i = \begin{cases} \infty & \text{if } y_{i1} + y_{i2} = 2, \\ -\infty & \text{if } y_{i1} + y_{i2} = 0, \\ -\dfrac{\beta}{2} & \text{if } y_{i1} + y_{i2} = 1. \end{cases} \tag{7.3.4}$$

Inserting (7.3.4) into (7.3.2), and letting n_1 denote the number of individuals

with $y_{i1} + y_{i2} = 1$ and n_2 the number of individuals with $y_{i1} + y_{i2} = 2$, we have[7]

$$\sum_{i=1}^{N} \frac{e^{\beta+\alpha_i}}{1+e^{\beta+\alpha_i}} = n_1 \frac{e^{\beta/2}}{1+e^{\beta/2}} + n_2 = \sum_{i=1}^{N} y_{i2}. \tag{7.3.5}$$

Therefore,

$$\hat{\beta} = 2\left\{\log\left(\sum_{i=1}^{N} y_{i2} - n_2\right) - \log\left(n_1 + n_2 - \sum_{i=1}^{N} y_{i2}\right)\right\}. \tag{7.3.6}$$

By a law of large numbers (Rao (1973, Chapter 2)),

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} \frac{1}{N}\left(\sum_{i=1}^{N} y_{i2} - n_2\right) &= \frac{1}{N}\sum_{i=1}^{N} \text{Prob}(y_{i1} = 0, y_{i2} = 1 \mid \beta, \alpha_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\beta+\alpha_i}}{(1+e^{\alpha_i})(1+e^{\beta+\alpha_i})}, \end{aligned} \tag{7.3.7}$$

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} \frac{1}{N}\left(n_1 + n_2 - \sum_{i=1}^{N} y_{i2}\right) &= \frac{1}{N}\sum_{i=1}^{N} \text{Prob}(y_{i1} = 1, y_{i2} = 0 \mid \beta, \alpha_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\alpha_i}}{(1+e^{\alpha_i})(1+e^{\beta+\alpha_i})}. \end{aligned} \tag{7.3.8}$$

Substituting $\hat{\alpha}_i = -\frac{\beta}{2}$ into (7.3.7) and (7.3.8), we obtain

$$\operatorname*{plim}_{N\to\infty} \hat{\beta} = 2\beta, \tag{7.3.9}$$

which is not consistent.

7.3.1.b Conditions for the Existence of a Consistent Estimator

Neyman and Scott (1948) suggested a general principle to find a consistent estimator for the (structural) parameter $\boldsymbol{\beta}$ in the presence of the incidental parameters α_i.[8] Their idea is to find K functions

$$\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \boldsymbol{\beta}), \qquad j = 1, \ldots, K, \tag{7.3.10}$$

that are independent of the incidental parameters α_i and have the property that when $\boldsymbol{\beta}$ are the true values, $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \boldsymbol{\beta})$ converges to zero in probability as N tends to infinity. Then an estimator $\hat{\boldsymbol{\beta}}$ derived by solving $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \hat{\boldsymbol{\beta}}) = 0$ is consistent under suitable regularity conditions. For instance, $\hat{\beta}^* = (\frac{1}{2})\hat{\beta}$ for the foregoing example of a fixed-effect logit model (7.3.1)–(7.3.3) is such an estimator.

In the case of a linear-probability model, either taking first differences over time or taking differences with respect to the individual mean eliminates the

individual-specific effect. The least-squares regression of the differenced equations yields a consistent estimator for $\boldsymbol{\beta}$ when N tends to infinity.

But in the general nonlinear models, simple functions for Ψ are not always easy to find. For instance, in general we do not know the probability limit of the MLE of a fixed-effects logit model. However, if a minimum sufficient statistic τ_i for the incidental parameter α_i exists and is not dependent on the structural parameter $\boldsymbol{\beta}$, then the conditional density,

$$f^*(\mathbf{y}_i \mid \boldsymbol{\beta}, \tau_i) = \frac{f(\mathbf{y}_i \mid \boldsymbol{\beta}, \alpha_i)}{g(\tau_i \mid \boldsymbol{\beta}, \alpha_i)} \qquad \text{for } g(\tau_i \mid \boldsymbol{\beta}, \alpha_i) > 0, \tag{7.3.11}$$

no longer depends on α_i.[9] Andersen (1970, 1973) has shown that maximizing the conditional density of $\mathbf{y}_1, \ldots, \mathbf{y}_N$ given $\tau_1, \ldots, \tau_N$,

$$\prod_{i=1}^{N} f^*\left(\mathbf{y}_i \mid \boldsymbol{\beta}, \tau_i\right), \tag{7.3.12}$$

yields the first-order conditions $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \hat{\boldsymbol{\beta}}, \tau_1, \tau_2, \ldots \tau_N) = 0$ for $j = 1, \ldots, K$. Solving for these functions will give a consistent estimator of the common (structural) parameter $\boldsymbol{\beta}$ under mild regularity conditions.[10]

To illustrate the conditional maximum likelihood method, we use the logit model as an example. The joint probability of $\mathbf{y}_i$ is

$$\text{Prob}(\mathbf{y}_i) = \frac{\exp\left\{\alpha_i \sum_{t=1}^{T} y_{it} + \boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\right\}}{\prod_{t=1}^{T}[1 + \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i)]}. \tag{7.3.13}$$

It is clear that $\sum_{t=1}^{T} y_{it}$ is a minimum sufficient statistic for α_i. The conditional probability for $\mathbf{y}_i$, given $\sum_{t=1}^{T} y_{it}$, is

$$\text{Prob}\left(\mathbf{y}_i \,\middle|\, \sum_{t=1}^{T} y_{it}\right) = \frac{\exp\left[\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\right]}{\sum_{D_i \in \tilde{B}_i} \exp\left\{\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it} d_{ijt}\right\}}, \tag{7.3.14}$$

where $\tilde{B}_i = \{D_{ij} = (d_{ij1}, \ldots, d_{ijT}) \mid d_{ijt} = 0 \text{ or } 1 \text{ and } \sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it} = s,\ j = 1, 2, \ldots, \frac{T!}{s!(T-s)!}\}$ is the set of all possible distinct sequences $(d_{ij1}, d_{ij2}, \ldots d_{ijT})$ satisfying $\sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it} = s$. There are $T + 1$ distinct alternative sets corresponding to $\sum_{t=1}^{T} y_{it} = 0, 1, \ldots, T$. Groups for which $\sum_{t=1}^{T} y_{it} = 0$ or T contribute zero to the likelihood function, because the corresponding conditional probability in this case is equal to 1 (with $\alpha_i = -\infty$ or ∞). So only $T - 1$ alternative sets are relevant. The alternative sets for groups with $\sum_{t=1}^{T} y_{it} = s$ have $\binom{T}{s}$ elements, corresponding to the distinct sequences of T trials with s successes.

Equation (7.3.14) is in a conditional logit form (McFadden (1974)), with the alternative sets $(\tilde{B}_i)$ varying across observations i. It does not depend on the incidental parameters α_i. Therefore, the conditional maximum likelihood estimator of $\boldsymbol{\beta}$ can be obtained by using standard maximum likelihood logit programs, and it is consistent under mild conditions. For example, with $T = 2$,

the only case of interest is $y_{i1} + y_{i2} = 1$. The two possibilities are $\omega_i = 1$, if $(y_{i1}, y_{i2}) = (0, 1)$, and $\omega_i = 0$, if $(y_{i1}, y_{i2}) = (1, 0)$.

The conditional probability of $w_i = 1$ given $y_{i1} + y_{i2} = 1$ is

$$\begin{aligned}\text{Prob}(\omega_i = 1 \mid y_{i1} + y_{i2} = 1) &= \frac{\text{Prob}(\omega_i = 1)}{\text{Prob}(\omega_i = 1) + \text{Prob}(\omega_i = 0)} \\ &= \frac{\exp[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]}{1 + \exp[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]} \\ &= F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]. \qquad (7.3.15)\end{aligned}$$

Equation (7.3.15) is in the form of a binary logit function in which the two outcomes are (0, 1) and (1, 0), with explanatory variables $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$. The conditional log likelihood function is

$$\begin{aligned}\log L^* = \sum_{i \in \tilde{B}_1} \{&\omega_i \log F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \\ &+ (1 - \omega_i) \log(1 - F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})])\}, \qquad (7.3.16)\end{aligned}$$

where $\tilde{B}_1 = \{i \mid y_{i1} + y_{i2} = 1\}$.

Although $\tilde{B}_1$ is a random set of indices, Chamberlain (1980) has shown that the inverse of the information matrix based on the conditional-likelihood function provides an asymptotic covariance matrix for the conditional MLE of $\boldsymbol{\beta}$ as N tends to infinity. This can be made more explicit by defining $d_i = 1$ if $y_{i1} + y_{i2} = 1$, and $d_i = 0$ otherwise, for the foregoing case in which $T = 2$. Then we have

$$\begin{aligned}J_{\tilde{B}_1} = \frac{\partial^2 \log L^*}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} &= -\sum_{i=1}^{N} d_i F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \\ &\times \{1 - F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]\}(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \cdot (\mathbf{x}_{i2} - \mathbf{x}_{i1})'. \qquad (7.3.17)\end{aligned}$$

The information matrix is

$$\begin{aligned}J &= E(J_{\tilde{B}_1}) \\ &= -\sum_{i=1}^{N} P_i F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \\ &\quad \times \{1 - F[\boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]\}(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \cdot (\mathbf{x}_{i2} - \mathbf{x}_{i1})', \qquad (7.3.18)\end{aligned}$$

where $P_i = E(d_i \mid \alpha_i) = F(\boldsymbol{\beta}'\mathbf{x}_{i1} + \alpha_i)[1 - F(\boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i)] + [1 - F(\boldsymbol{\beta}'\mathbf{x}_{i1} + \alpha_i)]F(\boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i)$. Because d_i are independent, with $Ed_i = P_i$, and both F and the variance of d_i are uniformly bounded, by a strong law of large numbers we have

$$\frac{1}{N} J_{\tilde{B}_1} - \frac{1}{N} J \to 0 \qquad \text{almost surely as } N \to \infty$$

$$\text{if} \quad \sum_{i=1}^{N} \frac{1}{i^2} \mathbf{m}_i \mathbf{m}_i' < \infty, \qquad (7.3.19)$$

where $\mathbf{m}_i$ replaces each element of $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ with its square. The condition for convergence clearly holds if $\mathbf{x}_{it}$ is uniformly bounded.

For the case of $T > 2$, there is no loss of generality in choosing the sequence $D_{i1} = (d_{i11}, \ldots, d_{i1T})$, $\sum_{t=1}^{T} d_{i1t} = \sum_{t=1}^{T} y_{it} = s$, $1 \leq s \leq T-1$, as the normalizing factor. Hence we may rewrite the conditional probability (7.3.14) as

$$\text{Prob}\left(\mathbf{y}_i \mid \sum_{t=1}^{T} y_{it}\right) = \frac{\exp\left\{\boldsymbol{\beta}'\left\{\sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t})\right\}\right\}}{1 + \sum_{D_{ij} \in (B_i - D_{i1})} \exp\left\{\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t})\right\}}. \tag{7.3.20}$$

Then the conditional log-likelihood function takes the form

$$\log L^* = \sum_{i \in C} \left\{ \boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t}) - \log\left[1 + \sum_{D_{ij} \in (B_i - D_{i1})} \exp\left\{\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t})\right\}\right]\right\}, \tag{7.3.21}$$

where $C = \{i \mid \sum_{t=1}^{T} y_{it} \neq T, \sum_{t=1}^{T} y_{it} \neq 0\}$.

Although we can find simple transformations of linear-probability and logit models that will satisfy the Neyman–Scott principle, we cannot find simple functions for the parameters of interest that are independent of the nuisance parameters α_i for probit models. That is, there does not appear to exist a consistent estimator of $\boldsymbol{\beta}$ for the fixed-effects probit models.

7.3.1.c Some Monte Carlo Evidence

Given that there exists a consistent estimator of $\boldsymbol{\beta}$ for the fixed-effects logit model, but not for the fixed-effects probit model, and that in the binary case the probit and logit models yield similar results, it appears that a case can be made for favoring the logit specification because of the existence of a consistent estimator for the structural parameter $\boldsymbol{\beta}$. However, in the multivariate case, logit and probit models yield very different results. In this situation it will be useful to know the magnitude of the bias if the data actually call for a fixed-effects probit specification.

Heckman (1981b) conducted a limited set of Monte Carlo experiments to get some idea of the order of bias of the MLE for the fixed-effects probit models. His data were generated by the model

$$y_{it}^* = \beta x_{it} + \alpha_i + u_{it}, \qquad i = 1, 2, \ldots, N, \quad t = 1, \ldots, T, \tag{7.3.22}$$

Table 7.1. *Average values of $\hat{\beta}$ for the fixed-effects probit model*

	$\hat{\beta}$		
σ_α^2	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
3	0.90	−0.10	−0.94
1	0.91	−0.09	−0.95
0.5	0.93	−0.10	−0.96

Source: Heckman (1981b, Table 4.1).

and

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The exogenous variable x_{it} was generated by a Nerlove (1971a) process,

$$x_{it} = 0.1t + 0.5x_{i,t-1} + \epsilon_{it}, \tag{7.3.23}$$

where ϵ_{it} is a uniform random variable having mean zero and range $-1/2$ to $1/2$. The variance σ_u^2 was set at 1. The scale of the variation of the fixed effect, σ_α^2, is changed for different experiments. In each experiment, 25 samples of 100 individuals ($N = 100$) were selected for eight periods ($T = 8$).

The results of Heckman's experiment with the fixed-effects MLE of probit models are presented in Table 7.1. For $\beta = -0.1$, the fixed-effects estimator does well. The estimated value comes very close to the true value. For $\beta = -1$ or $\beta = 1$, the estimator does not perform as well, but the bias is never more than 10 percent and is always toward zero. Also, as the scale of the variation in the fixed-effects decreases, so does the bias.[11]

7.3.2 Random-Effects Models

When the individual specific effects α_i are treated as random, we may still use the fixed effects estimators to estimate the structural parameters $\boldsymbol{\beta}$. The asymptotic properties of the fixed effects estimators of $\boldsymbol{\beta}$ remain unchanged. However, if α_i are random, but are treated as fixed, the consequence, at its best, is a loss of efficiency in estimating $\boldsymbol{\beta}$, but it could be worse, namely, the resulting fixed-effects estimators may be inconsistent, as discussed in Section 7.3.1.

When α_i are independent of $\mathbf{x}_i$ and are a random sample from a univariate distribution G, indexed by a finite number of parameters $\boldsymbol{\delta}$, the log likelihood function becomes

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha)^{y_{it}} [1 - F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha)]^{1-y_{it}} \, dG(\alpha \mid \boldsymbol{\delta}), \tag{7.3.24}$$

where $F(\cdot)$ is the distribution of the error term conditional on both $\mathbf{x}_i$ and α_i. Equation (7.3.24) replaces the probability function for $\mathbf{y}$ conditional on α by a probability function that is marginal on α. It is a function of a finite number of parameters $(\boldsymbol{\beta}', \boldsymbol{\delta}')$. Thus, maximizing (7.3.24), under weak regularity conditions, will give consistent estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ as N tends to infinity.

If α_i is correlated with $\mathbf{x}_{it}$, maximizing (7.3.24) will not eliminate the omitted-variable bias. To allow for dependence between α and $\mathbf{x}$, we must specify a distribution $G(\alpha \mid \mathbf{x})$ for α conditional on $\mathbf{x}$, and consider the marginal log likelihood function

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha)^{y_{it}} \times [1 - F(\boldsymbol{\beta}'\mathbf{x}_{it} + \alpha)]^{1-y_{it}}\, dG(\alpha \mid \mathbf{x}). \tag{7.3.24'}$$

A convenient specification suggested by Chamberlain (1980, 1984) is to assume that $\alpha_i = \sum_{t=1}^{T} \mathbf{a}_t'\mathbf{x}_{it} + \eta_i = \mathbf{a}'\mathbf{x}_i + \eta_i$, where $\mathbf{a}' = (\mathbf{a}_1', \ldots \mathbf{a}_T')$, $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, and η_i is the residual. However, there is a very important difference in this step compared with the linear case. In the linear case it was not restrictive to decompose α_i into its linear projection on $\mathbf{x}_i$ and an orthogonal residual. Now we are assuming that the regression function $E(\alpha_i \mid \mathbf{x}_i)$ is actually linear, that η_i is independent of $\mathbf{x}_i$, and that η_i has a specific probability distribution.

Given these assumptions, the log likelihood function under our random-effects specification is

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\boldsymbol{\beta}'\mathbf{x}_{it} + \mathbf{a}'\mathbf{x}_i + \eta)^{y_{it}} \times [1 - F(\boldsymbol{\beta}'\mathbf{x}_{it} + \mathbf{a}'\mathbf{x}_i + \eta)]^{1-y_{it}}\, dG^*(\eta), \tag{7.3.25}$$

where G^* is a univariate distribution function for η. For example, if F is a standard normal distribution function and we choose G^* to be the distribution function of a normal random variable with mean 0 and variance σ_η^2, then our specification gives a multivariate probit model:

$$y_{it} = 1 \quad \text{if} \quad \boldsymbol{\beta}'\mathbf{x}_{it} + \mathbf{a}'\mathbf{x}_i + \eta_i + u_{it} > 0, \tag{7.3.26}$$

where $\mathbf{u}_i + \mathbf{e}\eta_i$ is independent normal, with mean $\mathbf{0}$ and variance–covariance matrix $I_T + \sigma_\eta^2 \mathbf{e}\mathbf{e}'$.

The difference between (7.3.25) and (7.3.24) is only in the inclusion of the term $\mathbf{a}'\mathbf{x}_i$ to capture the dependence between the incidental parameters α_i and $\mathbf{x}_i$. Therefore, the essential characteristics with regard to estimation of (7.3.24) and (7.3.25) are the same. So we shall discuss only the procedure to estimate the more general model (7.3.25).

Maximizing (7.3.25) involves integration of T dimensions, which can be computationally cumbersome. An alternative approach, which simplifies the

computation of the MLE to a univariate integration is to note that conditional on α_i, the error terms $v_{it} = \alpha_i + u_{it}$ are independently normally distributed with mean α_i and variance 1, with probability density denoted by $\phi(v_{it} \mid \alpha_i)$ (Heckman (1981a)). Then

$$\Pr(y_{i1}, \ldots, y_{iT}) = \int_{c_{i1}}^{b_{i1}} \cdots \int_{c_{iT}}^{b_{iT}} \prod_{t=1}^{T} \phi(v_{it} \mid \alpha_i) G(\alpha_i \mid \mathbf{x}_i)\, d\alpha_i\, dv_{i1} \cdots dv_{iT}$$

$$= \int_{-\infty}^{\infty} G(\alpha_i \mid \mathbf{x}_i) \prod_{t=1}^{T} [\Phi(b_{it} \mid \alpha_i) - \Phi(c_{it} \mid \alpha_i)]\, d\alpha_i, \qquad (7.3.27)$$

where $\Phi(\cdot \mid \alpha_i)$ is the cumulative distribution function (cdf) of $\phi(\cdot \mid \alpha_i)$, $c_{it} = -\boldsymbol{\beta}'\mathbf{x}_{it}$, $b_{it} = \infty$ if $y_{it} = 1$ and $c_{it} = -\infty$, $b_{it} = -\boldsymbol{\beta}'\mathbf{x}_{it}$ if $y_{it} = 0$, and $G(\alpha_i \mid \mathbf{x}_i)$ is the probability density function of α_i given $\mathbf{x}_i$. If $G(\alpha_i \mid \mathbf{x}_i)$ is assumed to be normally distributed with variance σ_α^2, the expression (7.3.27) reduces a T-dimensional integration to a single integral whose integrand is a product of one normal density and T differences of normal cdfs for which highly accurate approximations are available. For instance, Butler and Moffit (1982) suggest using Gaussian quadrature to achieve gains in computational efficiency. The Gaussian quadrature formula for evaluation of the necessary integral is the Hermite integration formula $\int_{-\infty}^{\infty} e^{-z^2} g(z)\, dz = \sum_{j=1}^{l} w_j g(z_j)$, where l is the number of evaluation points, w_j is the weight given to the jth point, and $g(z_j)$ is $g(z)$ evaluated at the jth point of z. The points and weights are available from Abramowitz and Stegun (1965) and Stroud and Secrest (1966).

A key question for computational feasibility of the Hermite formula is the number of points at which the integrand must be evaluated for accurate approximation. Several evaluations of the integral using four periods of arbitrary values of the data and coefficients on right-hand-side variables by Butler and Moffitt (1982) show that even two-point integration is highly accurate. Of course, in the context of a maximization algorithm, accuracy could be increased by raising the number of evaluation points as the likelihood function approaches its optimum.

Although maximizing (7.3.25) or (7.3.24) provides a consistent and efficient estimator for $\boldsymbol{\beta}$, computationally it is still fairly involved. However, if both u_{it} and η_i (or α_i) are normally distributed, a computationally simple approach that avoids numerical integration is to make use of the fact that the distribution for y_{it} conditional on $\mathbf{x}_i$ but marginal on α_i also has a probit form:

$$\text{Prob}(y_{it} = 1) = \Phi\left[\left(1 + \sigma_\eta^2\right)^{-1/2} (\boldsymbol{\beta}'\mathbf{x}_{it} + \mathbf{a}'\mathbf{x}_i)\right]. \qquad (7.3.28)$$

Estimating each of t cross-sectional univariate probit specifications by maximum likelihood gives $\hat{\boldsymbol{\pi}}_t$, $t = 1, 2, \ldots, T$, which will converge to[12]

$$\Pi = \left(1 + \sigma_\eta^2\right)^{-1/2} (I_T \otimes \boldsymbol{\beta}' + \mathbf{e}\mathbf{a}') \qquad (7.3.29)$$

as N tends to infinity. Therefore, consistent estimators of $(1 + \sigma_\eta^2)^{-1/2}\boldsymbol{\beta}$ and $(1 + \sigma_\eta^2)^{-1/2}\mathbf{a}$ can be easily derived from (7.3.29). We can then follow Heckman's suggestion (1981a) by substituting these estimated values

into (7.3.25) and optimizing the functions with respect to σ_η^2 conditional on $(1+\sigma_\eta^2)^{-1/2}\boldsymbol{\beta}$ and $(1+\sigma_\eta^2)^{-1/2}\mathbf{a}$.

A more efficient estimator that also avoids numerical integration is to impose the restriction (7.3.29) by $\boldsymbol{\pi} = \text{vec}(\boldsymbol{\Pi}') = \mathbf{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{a}', \sigma_\eta^2)$, and use a minimum-distance estimator (see Section 3.9), just as in the linear case. Chamberlain (1984) suggests that we choose $\hat{\boldsymbol{\theta}}$ to minimize[13]

$$[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})], \tag{7.3.30}$$

where $\hat{\Omega}$ is a consistent estimator of

$$\Omega = J^{-1}\Delta J^{-1}, \tag{7.3.31}$$

where

$$J = \begin{bmatrix} J_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & J_2 & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & J_T \end{bmatrix},$$

$$J_t = E\left\{\frac{\phi_{it}^2}{\Phi_{it}(1-\Phi_{it})}\mathbf{x}_i\mathbf{x}_i'\right\},$$

$$\Delta = E[\Psi_i \otimes \mathbf{x}_i\mathbf{x}_i'],$$

and where the t, s element of the $T \times T$ matrix Ψ_i is $\psi_{it} = c_{it}c_{is}$, with

$$c_{it} = \frac{y_{it} - \Phi_{it}}{\Phi_{it}(1-\Phi_{it})}\phi_{it}, \qquad t = 1, \ldots, T.$$

The standard normal distribution function Φ_{it} and the standard normal density function ϕ_{it} are evaluated at $\boldsymbol{\pi}'\mathbf{x}_i$. We can obtain a consistent estimator of Ω by replacing expectations by sample means and using $\hat{\boldsymbol{\pi}}$ in place of $\boldsymbol{\pi}$.

7.4 SEMIPARAMETRIC APPROACH TO STATIC MODELS

The parametric approach to estimating discrete choice models suffers from two drawbacks: (1) Conditional on $\mathbf{x}$, the probability law of generating (u_{it}, α_i) is known a priori, or conditional on $\mathbf{x}$ and α_i, the probability law of u_{it} is known a priori. (2) When α_i are fixed, it appears that apart from the logit and linear probability models, there does not exist a simple transformation that can get rid of the incidental parameters. The semiparametric approach not only avoids assuming a specific distribution of u_{it}, but also allows consistent estimation of $\boldsymbol{\beta}$ up to a scale, whether α_i is treated as fixed or random.

7.4.1 Maximum Score Estimator

Manski (1975, 1985, 1987) suggests a maximum score estimator that maximizes the sample average function

$$H_N(\mathbf{b}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T} \operatorname{sgn}(\Delta \mathbf{x}_{it}'\mathbf{b})\Delta y_{it} \tag{7.4.1}$$

subject to the normalization condition $\mathbf{b}'\mathbf{b} = 1$, where $\Delta \mathbf{x}_{it} = \mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, $\Delta y_{it} = y_{it} - y_{i,t-1}$, and $\operatorname{sgn}(w) = 1$ if $w > 0$, 0 if $w = 0$, and -1 if $w < 0$. This is because under fairly general conditions (7.4.1) converges uniformly to

$$H(\mathbf{b}) = E[\operatorname{sgn}(\Delta \mathbf{x}_{it}'\mathbf{b})\Delta y_{it}], \tag{7.4.2}$$

where $H(\mathbf{b})$ is maximized at $\mathbf{b} = \boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$ and $\|\boldsymbol{\beta}\|$ the Euclidean norm $\sum_{k=1}^{K}\beta_k^2$.

To see this, we note that the binary-choice model can be written in the form

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0, \\ 0 & \text{if } y_{it}^* \le 0, \end{cases} \tag{7.4.3}$$

where y_{it}^* is given by (7.2.1) with $v_{it} = \alpha_i + u_{it}$. Under the assumption that u_{it} is independently, identically distributed and is independent of $\mathbf{x}_i$ and α_i for given i, (i.e., $\mathbf{x}_{it}$ is strictly exogenous), we have

$$\begin{aligned} \mathbf{x}_{it}'\boldsymbol{\beta} > \mathbf{x}_{i,t-1}'\boldsymbol{\beta} \quad &\Leftrightarrow \quad E(y_{it} \mid \mathbf{x}_{it}) > E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}), \\ \mathbf{x}_{it}'\boldsymbol{\beta} = \mathbf{x}_{i,t-1}'\boldsymbol{\beta} \quad &\Leftrightarrow \quad E(y_{it} \mid \mathbf{x}_{it}) = E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}), \\ \mathbf{x}_{it}'\boldsymbol{\beta} < \mathbf{x}_{i,t-1}'\boldsymbol{\beta} \quad &\Leftrightarrow \quad E(y_{it} \mid \mathbf{x}_{it}) < E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}). \end{aligned} \tag{7.4.4}$$

Rewriting (7.4.4) in terms of first differences, we have the equivalent representation

$$\begin{aligned} \Delta \mathbf{x}_{it}'\boldsymbol{\beta} > 0 \quad &\Leftrightarrow \quad E(y_{it} - y_{i,t-1} \mid \Delta \mathbf{x}_{it}) > 0, \\ \Delta \mathbf{x}_{it}'\boldsymbol{\beta} = 0 \quad &\Leftrightarrow \quad E(y_{it} - y_{i,t-1} \mid \Delta \mathbf{x}_{it}) = 0, \\ \Delta \mathbf{x}_{it}'\boldsymbol{\beta} < 0 \quad &\Leftrightarrow \quad E(y_{it} - y_{i,t-1} \mid \Delta \mathbf{x}_{it}) < 0. \end{aligned} \tag{7.4.5}$$

It is obvious that (7.4.5) continues to hold for any $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}c$ where $c > 0$. Therefore, we shall only consider the normalized vector $\boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$.

Then, for any $\mathbf{b}$ (satisfying $\mathbf{b}'\mathbf{b} = 1$) such that $\mathbf{b} \neq \boldsymbol{\beta}^*$,

$$\begin{aligned} H(\boldsymbol{\beta}^*) - H(\mathbf{b}) &= E\{[\operatorname{sgn}(\Delta \mathbf{x}_{it}'\boldsymbol{\beta}^*) - \operatorname{sgn}(\Delta \mathbf{x}_{it}'\mathbf{b})](y_{it} - y_{i,t-1})\} \\ &= 2\int_{W_b} \operatorname{sgn}(\Delta \mathbf{x}_{it}'\boldsymbol{\beta}^*)E[y_t - y_{t-1} \mid \Delta \mathbf{x}]\, dF_{\Delta \mathbf{x}}, \end{aligned} \tag{7.4.6}$$

where $W_b = [\Delta \mathbf{x} : \operatorname{sgn}(\Delta \mathbf{x}'\boldsymbol{\beta}^*) \neq \operatorname{sgn}(\Delta \mathbf{x}'\mathbf{b})]$, and $F_{\Delta \mathbf{x}}$ denotes the distribution of $\Delta \mathbf{x}$. Because of (7.4.5) the relation (7.4.6) implies that for all $\Delta \mathbf{x}$,

$$\operatorname{sgn}(\Delta \mathbf{x}'\boldsymbol{\beta}^*)E[y_t - y_{t-1} \mid \Delta \mathbf{x}] = |E[y_t - y_{t-1} \mid \Delta \mathbf{x}]|.$$

Therefore, under the assumption on the $\mathbf{x}$'s,

$$H(\boldsymbol{\beta}^*) - H(\mathbf{b}) = 2\int_{W_b} |E[y_t - y_{t-1} \mid \Delta\mathbf{x}]| \, dF_{\Delta\mathbf{x}} > 0. \tag{7.4.7}$$

Manski (1985, 1987) has shown that under fairly general conditions, the estimator maximizing the criterion function (7.4.1) yields a strongly consistent estimator of $\boldsymbol{\beta}^*$.

As discussed in Chapter 3 and early sections of this chapter, when T is small the MLE of the (structural) parameters $\boldsymbol{\beta}$ is consistent as $N \to \infty$ for the linear model and inconsistent for the nonlinear model in the presence of incidental parameters α_i, because in the former case we can eliminate α_i by differencing, while in the latter case we cannot. Thus, the error of estimating α_i is transmitted into the estimator of $\boldsymbol{\beta}$ in the nonlinear case. The semiparametric approach allows one to make use of the linear structure of the latent-variable representation (7.2.1) or (7.4.4). The individual-specific effects α_i can again be eliminated by differencing, and hence the lack of knowledge of α_i no longer affects the estimation of $\boldsymbol{\beta}$.

The Manski maximum score estimator is consistent as $N \to \infty$ if the conditional distribution of u_{it} given α_i and $\mathbf{x}_{it}, \mathbf{x}_{i,t-1}$ is identical to the conditional distribution of $u_{i,t-1}$ given α_i and $\mathbf{x}_{it}, \mathbf{x}_{i,t-1}$. However, it converges at the rate $N^{1/3}$, which is much slower than the usual speed of $N^{1/2}$ for the parametric approach. Moreover, Kim and Pollard (1990) have shown that $N^{1/3}$ times the centered maximum score estimator converges in distribution to the random variable that maximizes a certain Gaussian process. This result cannot be used in application, since the properties of the limiting distribution are largely unknown.

The objective function (7.4.1) is equivalent to

$$\max_{\mathbf{b}} H_N^*(\mathbf{b}) = N^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}[2 \cdot \mathbf{1}(\Delta y_{it} = 1) - 1]\mathbf{1}(\Delta\mathbf{x}_{it}'\mathbf{b} \geq \mathbf{0}), \tag{7.4.8}$$

subject to $\mathbf{b}'\mathbf{b} = 1$, where $\mathbf{1}(A)$ is the indicator of the event A, with $\mathbf{1}(A) = 1$ if A occurs and 0 otherwise. The complexity of the maximum score estimator and its slow convergence are due to the discontinuity of the function $H_N(\mathbf{b})$ or $H_N^*(\mathbf{b})$. Horowitz (1992) suggests avoiding these difficulties by replacing $H_N^*(\mathbf{b})$ with a sufficiently smooth function $\tilde{H}_N(\mathbf{b})$ whose almost sure limit as $N \to \infty$ is the same as that of $H_N^*(\mathbf{b})$. Let $K(\cdot)$ be a continuous function of the real line into itself such that

i. $|K(v)| < M$ for some finite M and all v in $(-\infty, \infty)$,
ii. $\lim_{v\to-\infty} K(v) = 0$ and $\lim_{v\to\infty} K(v) = 1$.

The $K(\cdot)$ here is analogous to a cumulative distribution function. Let $\{\sigma_N : N = 1, 2, \ldots\}$ be a sequence of strictly positive real numbers satisfying $\lim_{N\to\infty} \sigma_N = 0$. Define

$$\tilde{H}_N(\mathbf{b}) = N^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}[2 \cdot \mathbf{1}(\Delta y_{it} = 1) - 1]K(\mathbf{b}'\Delta\mathbf{x}_{it}/\sigma_N). \tag{7.4.9}$$

Horowitz (1992) defines a smoothed maximum score estimator as any solution that maximizes (7.4.9). Like Manski's estimator, $\boldsymbol{\beta}$ can be identified only up to scale. Instead of using the normalization $\|\boldsymbol{\beta}^*\| = 1$, Horowitz (1992) finds it more convenient to use the normalization that the coefficient of one component of $\Delta\mathbf{x}$, say Δx_1, is to be equal to 1 in absolute value if $\beta_1 \neq 0$, and the probability distribution of Δx_1 conditional on the remaining components is absolutely continuous (with respect to Lebesgue measure).

The smoothed maximum score estimator is strongly consistent under the assumption that the distribution of $\Delta u_{it} = u_{it} - u_{i,t-1}$ conditional on $\Delta\mathbf{x}_{it}$ is symmetrically distributed with mean equal to zero. The asymptotic behavior of the estimator can be analyzed using the Taylor series methods of asymptotic theory by taking a Taylor expansion of the first-order conditions and applying a version of the central limit theorem and the law of large numbers. The smoothed estimator of $\boldsymbol{\beta}$ is consistent and, after centering and suitable normalization, is asymptotically normally distributed. Its rate of convergence is at least as fast as $N^{-2/5}$ and, depending on how smooth the distribution of u and $\boldsymbol{\beta}'\Delta\mathbf{x}$ are, can be arbitrarily close to $N^{-1/2}$.

7.4.2 A Root-N Consistent Semiparametric Estimator

The speed of convergence of the smoothed maximum score estimator depends on the speed of convergence of $\sigma_N \to 0$. Lee (1999) suggests a root-N consistent semiparametric estimator that does not depend on a smoothing parameter by maximizing the double sums

$$\{N(N-1)\}^{-1} \sum_{i \neq j} \sum_{t=2}^{T} \operatorname{sgn}(\Delta\mathbf{x}_{it}'\mathbf{b} - \Delta\mathbf{x}_{jt}'\mathbf{b})(\Delta y_{it} - \Delta y_{jt})\Delta y_{it}^2\, \Delta y_{jt}^2$$

$$= \binom{N}{2}^{-1} \sum_{\substack{i \\ i<j,\ \Delta y_{it} \neq \Delta y_{jt}}} \sum_{\substack{j \\ \Delta y_{it} \neq 0,\ \Delta y_{jt} \neq 0}} \sum_{t=2}^{T} \operatorname{sgn}(\Delta\mathbf{x}_{it}'\mathbf{b} - \Delta\mathbf{x}_{jt}'\mathbf{b})(\Delta y_{it} - \Delta y_{jt}) \tag{7.4.10}$$

with respect to $\mathbf{b}$. The consistency of the Lee estimator $\hat{\mathbf{b}}$ follows from the fact that although $\Delta y_{it} - \Delta y_{jt}$ can take five values $(0, \pm 1, \pm 2)$, the event that $(\Delta y_{it} - \Delta y_{jt})\Delta y_{it}^2\, \Delta y_{jt}^2 \neq 0$ excludes $(0, \pm 1)$ and thus makes $\Delta y_{it} - \Delta y_{jt}$ binary (2 or -2). Conditional on given j, the first average over i and t converges to

$$E\left\{\operatorname{sgn}(\Delta\mathbf{x}'\mathbf{b} - \Delta\mathbf{x}_j'\mathbf{b})(\Delta y - \Delta y_j)\Delta y^2\, \Delta y_j^2 \mid \Delta\mathbf{x}_j, \Delta y_j\right\}. \tag{7.4.11}$$

The $\sqrt{N}$ speed of convergence follows from the second average of the smooth function (7.4.10).

Normalizing $\beta_1 = 1$, the asymptotic covariance matrix of $\sqrt{N}(\tilde{\mathbf{b}} - \tilde{\boldsymbol{\beta}})$ is equal to

$$4 \cdot (E\nabla_2\boldsymbol{\tau})^{-1}(E\nabla_1\boldsymbol{\tau}\nabla_1\boldsymbol{\tau}')(E\nabla_2\boldsymbol{\tau})^{-1}, \tag{7.4.12}$$