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 is only on controlling for unobserved characteristics of individual units to avoid specification bias. In Section 7.2, we briefly review some popular parametric specifications for cross-sectional data. Sections 7.3 and 7.4 discuss inference of panel parametric and semiparametric static models with heterogeneity, respectively. Section 7.5 discusses dynamic models. Section 7.6 discusses alternative approaches to identify state dependence.

7.2 SOME DISCRETE-RESPONSE MODELS FOR CROSS-SECTIONAL DATA

In this section we briefly review some widely used discrete-response models for cross-sectional data. We suppose there are observations for $K + 1$ variables $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, where the dependent variable y_i can take only two values, which for convenience and without any loss of generality will be the value of 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_i can be viewed as the observed counterpart of a latent continuous random variable crossing a threshold. Suppose that the continuous latent random variable, y_i^*, is a linear function of a vector of explanatory variable, $\mathbf{x}_i$,

$$y_i^* = \beta' \mathbf{x}_i + v_i, \tag{7.2.1}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

where the error term v_i is independent of $\mathbf{x}_i$ with mean 0. Suppose, instead of observing y_i^*, we observe y_i, where

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

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

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

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

$$y_i = \beta'\mathbf{x}_i + v_i \tag{7.2.4}$$

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

$$\begin{aligned} Pr(y_i = 1 \mid \mathbf{x}_i) &= \int_{-\beta'\mathbf{x}_i}^{\infty} \phi(v_i) dv_i \\ &= \int_{-\infty}^{\beta'\mathbf{x}_i} \phi(v_i) dv_i = \Phi(\beta'\mathbf{x}_i). \end{aligned} \tag{7.2.5}$$

When the probability density function is a standard logistic,

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

we have the logit model

$$Pr(y_i = 1 \mid \mathbf{x}_i) = \int_{-\beta'\mathbf{x}_i}^{\infty} \frac{\exp(v_i)}{(1 + \exp(v_i))^2} dv_i = \frac{\exp(\beta'\mathbf{x}_i)}{1 + \exp(\beta'\mathbf{x}_i)}. \tag{7.2.6}$$

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

Linear-Probability Model

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

Probit model

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv = \Phi(w) \tag{7.2.8}$$

Logit model

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

The linear-probability model is a special case of the linear regression model with heteroscadastic variance, $\beta'\mathbf{x}_i(1-\beta'\mathbf{x}_i)$. It can be estimated by least-squares or weighted least-squares (Goldberger 1964). But it has an obvious defect in that $\beta'\mathbf{x}_i$ is not constrained to lie between 0 and 1 as a probability should, whereas the probit and logit models do. 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, the probability an event occurs depends only on $(\frac{1}{\sigma})\beta'\mathbf{x}$, where σ denotes the standard deviation of a normal density. 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, except that 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 three models. Let the coefficients for the linear-probability, probit, and logit models be denoted as $\hat{\beta}_{LP}, \hat{\beta}_{\Phi}, \hat{\beta}_{L}$, respectively. Then

$$\hat{\beta}_L \simeq 1.6\,\hat{\beta}_{\Phi},$$

$$\hat{\beta}_{LP} \simeq 0.4\,\hat{\beta}_{\Phi} \text{ except for the constant term,} \tag{7.2.10}$$

and

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

For random sample of N individuals, the likelihood function for these three models can be written in general form as

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

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

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

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

where $F'(\beta'\mathbf{x}_i)$ and $F''(\beta'\mathbf{x}_i)$ denote the first and second derivatives of $F(\beta'\mathbf{x}_i)$ with respect to $\beta'\mathbf{x}_i$. If the likelihood function (7.2.11) is concave, as in the models discussed here (e.g., Amemiya 1985, p. 273), a Newton–Raphson method,

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

or a method of scoring,

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

can be used to find the maximum-likelihood estimator of β, where $\hat{\beta}^{(j)}$ 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 $p = F(\beta'\mathbf{x})$. Taking the inverse of this function yields $F^{-1}(p) = \beta'\mathbf{x}$. Substituting $\hat{p}$ for p, we have $F^{-1}(\hat{p}) = \beta'\mathbf{x} + \zeta$, 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 β. The resulting estimator, which is generally referred to as the minimum-chi-square estimator, has the same asymptotic efficiency as the maximum-likelihood estimator (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, 1980a; 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 for that value of $\mathbf{x}$ is not defined, but those observations can still be utilized to obtain the MLE. 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, say y_i takes $m_i + 1$ possible values, $0, 1, \ldots, m_i$, we can introduce $(m_i + 1)$ binary variables with

$$y_{ij} = \begin{cases} 1, & \text{if } y_i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{7.2.16}$$

$$i = 1, \ldots, N, \quad j = 0, 1, \ldots, m_i.$$

[1] For a survey of the minimum-chi-square method, see Hsiao (1985b).

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Let Prob $(y_i = j) = \text{Prob}\ (y_{ij} = 1) = F_{ij}$. If the sample is randomly drawn, the likelihood function takes the form

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

which is similar to the binary case (7.2.11). The complication is in the specification of F_{ij}. Once F_{ij} is specified, general results concerning the methods of estimation and inference of the dichotomous case also apply here.

If there is a natural ordering of the outcomes, say

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

one can use a single latent response function

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + v_i \tag{7.2.18}$$

to characterize the ordered outcomes with

$$y_i = \begin{cases} 0 & \text{if} \quad y_i^* < c_1, \\ 1 & \text{if} \quad c_1 < y_i^* < c_2, \\ 2 & \text{if} \quad c_2 < y_i^*. \end{cases} \tag{7.2.19}$$

If the outcomes are unordered, for instance,

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

then we will have to use a multivariate probability distribution to characterize the outcomes. One way to postulate unordered outcomes is to assume that the jth alternative is chosen because it yields higher utility than the utility of other alternatives. Let the ith individual's utility of choosing jth alternative be

$$y_{ij}^* = \mathbf{x}_i' \beta_j + v_{ij},\ j = 0, 1, \ldots, m_i. \tag{7.2.20}$$

Then

$$\begin{aligned} \text{Prob}(y_i = j \mid \mathbf{x}_i) &= \text{Prob}(y_{ij}^* > y_{i\ell}^*, \quad \forall\, \ell \neq j \mid \mathbf{x}_i) \\ &= F_{ij}. \end{aligned} \tag{7.2.21}$$

The probability F_{ij} is derived from the joint distribution of $(v_{io}, \ldots, v_{im})$. If $(v_{io}, \ldots, v_{im})$ follows a multivariate normal distribution, then (7.2.21) yields a multivariate probit. If the errors v_{ij} are independently, identically distributed with type I extreme value distribution, (7.2.21) yields a conditional logit model (McFadden 1974). 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 advantage of multivariate probit model is that it allows the choice among alternatives to have arbitrary correlation. The disadvantage is that the evaluation of Prob$(y_i = j)$ involves multiple

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

integrations that can be computationally infeasible. The advantage of the conditional logit model is that the evaluation of Prob($y_i = j$) does not involve multiple integration. The 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), Small and Hsiao (1985).

Because in many cases, a multi-response 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 only on the dichotomous model.[3]

When there is no information about the probability laws of generating v_i, a semi-parametric approach can be used to estimate β subject to certain normalization rule (e.g., Klein and Spady 1993; Manski 1985; Powell, Stock, and Stoker 1989). However, whether an investigator takes a parametric or semi-parametric 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 to distinguish average behavior from individual behavior by decomposing the error term, v_{it}, into

$$v_{it} = \alpha_i + \lambda_t + u_{it} \tag{7.2.22}$$

where α_i and λ_t denote the effects of omitted individual-specific and time-specific variables, respectively. Then Prob($y_i = 1 \mid \mathbf{x}, \alpha_i$) $\neq$ Prob($y_j = 1 \mid \mathbf{x}, \alpha_j$) if $\alpha_i \neq \alpha_j$. In this chapter, we 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 sum of the individual-specific effect and the time-varying omitted-variable effect as a pure chance event. However, as the example in Chapter 1 shows, a discovery of a group of married women having an average yearly labor participation rate of 50 percent could lead to

[2] The variable y_{i0} is sometimes omitted from the specification because it is determined by $y_{i0} = 1 - \Sigma_{j=1}^{m} y_{ij}$. For instance, a dichotomous model is often simply characterized by a single binary variable $y_i, i = 1, \ldots, N$.

[3] It should be noted that in generalizing the results of the binary case to the multiresponse case, we should allow for the fact that although y_{ij} and $y_{i'j}$ are independent for $i \neq i'$, y_{ij} and $y_{ij'}$ are not, because $\text{Cov}(y_{ij}, y_{ij'}) = -F_{ij}F_{ij'}$.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

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 finding relying on given cross-sectional data. To discriminate among the many possible explanations, we need information on individual labor-force histories in different subintervals of life cycle. Panel data, having information on intertemporal dynamics of individual entities, provide the possibility to separate a model of individual behavior from a model of average behavior of a group of individuals.

Suppose there are sample observations $(y_{it}, \mathbf{x}_{it})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where y_{it} is binary with $y_{it} = 1$ if y_{it}^* given by (7.2.1) is greater than 0 and 0 otherwise. 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 α_i are treated as fixed, $\text{Var}(v_{it} \mid \alpha_i) = \text{Var}(u_{it}) = \sigma_u^2$. When α_i are treated as random, we assume that $E\alpha_i = E\alpha_i u_{it} = 0$, and $\text{Var}(v_{it}) = \sigma_u^2 + \sigma_\alpha^2$. 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 of u, σ_u^2, to be equal to 1 for the parametric specifications discussed in Section 7.2.

The existence of such unobserved permanent components allows individuals who are homogeneous in terms of their observed characteristics to be heterogeneous in response probabilities, $F(\beta'\mathbf{x}_{it} + \alpha_i)$. For example, heterogeneity will imply that the sequential-participation behavior of a woman, $F(\beta'\mathbf{x} + \alpha_i)$, within a group of women with observationally identical $\mathbf{x}$ differ systematically from $F(\beta'\mathbf{x})$ or the average behavior of the group, $\int F(\beta'\mathbf{x} + \alpha) dH(\alpha \mid \mathbf{x})$, 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 the statistical inference of the common parameters β based on a parametric specification of $F(\cdot)$.

7.3.1 Fixed-Effects Models

7.3.1.1 Maximum-Likelihood Estimator

If the individual specific effect, α_i, is assumed to be fixed,[6] then both α_i and β are unknown parameters to be estimated for the model $\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i) = F(\beta'\mathbf{x}_{it} + \alpha_i)$. When T is finite, there is only a limited number of observations to provide information on α_i. Thus, we have the familiar incidental-parameter problem (Neyman and Scott 1948). Any estimation of the α_i is meaningless

[4] For a random-coefficient formulation of probit models, see Hausman and Wise (1978).

[5] Note that, in general, $\int F(\boldsymbol{\beta}'\mathbf{x} + \alpha) dH(\alpha \mid \mathbf{x}) \neq F[\boldsymbol{\beta}'\mathbf{x} + E(\alpha \mid \mathbf{x})]$.

[6] Note that for notational ease, we now use only α_i instead of both α_i and α_i^*. Readers should bear in mind that whenever α_i are treated as fixed, they are not viewed as the deviation from the common mean μ; rather, they are viewed as the sum of μ and the individual deviation. On the other hand, when α_i are treated as random, we assume that $E\alpha_i = 0$.

if we intend to judge the estimators by their large-sample properties. We shall therefore concentrate on estimation of the common parameters, β.

Unfortunately, contrary to the linear-regression case where the individual effects α_i can be eliminated by taking a linear transformation such as first difference, in general, it is hard to find simple transformation to eliminate the incidental parameters from a nonlinear model. The MLEs for α_i and β 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 β. Hence, even if β is the same for all i and t the MLE of β remains inconsistent if T is finite no matter how large N is.

We demonstrate the inconsistency of the MLE for β by considering a logit model. The log-likelihood function for this model is

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

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

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

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

Solving (7.3.3), we have

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

Inserting (7.3.4) into (7.3.2) and letting n_1 denote the number of individuals with $y_{i1} + y_{i2} = 1$ and letting n_2 denote 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,

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

[7] The number of individuals with $y_{i1} + y_{i2} = 0$ is $N - n_1 + n_2$.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

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

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

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

Substituting $\hat{\alpha}_i = \frac{\beta}{2}$ into (7.3.7) and (7.3.8) yields

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

which is not consistent.

7.3.1.2 Conditions for the Existence of a Consistent Estimator

Neyman and Scott (1948) have suggested a general principle to find a consistent estimator for the (structural) parameter β in the presence of the incidental parameters α_i.[8] Suppose the dimension of β is K, their idea is to find K functions

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

that are independent of the incidental parameters α_i and have the property that when β are the true values $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \beta)$ converges to some known constant, say 0, in probability as N tends to infinity. Then an estimator $\hat{\beta}$ derived by solving $\Psi_{Nj}\left(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \hat{\beta}\right) = 0$ is consistent under suitable regularity conditions. For instance, $\hat{\beta}^* = (1/2)\hat{\beta}$ 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 difference over time or taking difference with respect to the individual mean eliminates the

[8] We call $\boldsymbol{\beta}$ the structural parameter because the value of $\boldsymbol{\beta}$ characterizes the structure of the complete sequence of random variables. It is the same for all i and t. We call α_i an incidental parameter to emphasize that the value of α_i changes when i changes.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

individual-specific effect. The least-squares regression of the differenced equations yields a consistent estimator for β when N tends to infinity.

But in the general nonlinear models, simple forms of $\Psi(\cdot)$ 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 β, the conditional density,

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

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 $\tau_1, \ldots, \tau_N$,

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

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

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

The logit form has the property that the denominator of Prob(y_{it}) is always $[1 + \exp (\boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i)]$ independent of whether $y_{it} = 1$ or 0. On the other hand, for any sequence of dummy variable d_{ijt}, $D_{ij} = (d_{ij1}, d_{ij2}, \ldots, d_{ijT})$ where $d_{ijt} = 0$ or 1, the numerator of Prob(D_{ij}) always has the form $\exp(\alpha_i \sum_{t=1}^{T} d_{ijt}) \cdot \exp [\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it} d_{ijt}]$. It is clear that $\sum_{t=1}^{T} y_{it}$ is a minimum sufficient statistic for α_i. The conditional probability for y_{it} given $\sum_{t=1}^{T} y_{it}$ is

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

where $\bar{B}_i = \{D_{ij} = (d_{ij1}, \ldots, d_{ijT}) \mid d_{ijt} = 0 \text{ or } 1, \text{ and } \sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it}, j = 1, 2, \ldots, \frac{T!}{s!(T-s)!}\}$, is the set of all possible distinct sequence

[9] Suppose that the observed random variables $\mathbf{y}$ have a certain joint distribution function that belongs to a specific family $\mathcal{J}$ of distribution functions. The statistic $S(\mathbf{y})$ (a function of the observed sample values $\mathbf{y}$) is called a sufficient statistic if the conditional expectation of any other statistic $H(\mathbf{y})$, given $S(\mathbf{y})$, is independent of $\mathcal{J}$. A statistic $S^*(\mathbf{y})$ is called a minimum sufficient statistic if it is a function of every sufficient statistic $S(\mathbf{y})$ for $\mathcal{J}$. For additional discussion, see Zacks (1971, Chapter 2).

[10] When u_{it} are independently normally distributed, the LSDV estimator of $\boldsymbol{\beta}$ for the linear static model is the conditional MLE (Cornwell and Schmidt 1984).

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

$(d_{ij1}, d_{ij2}, \ldots, d_{ijT})$ 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 0 to the likelihood function, because the corresponding 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 success.

Equation (7.3.14) is in a conditional logit form (McFadden 1974), with the alternative sets $(\bar{B}_i)$ varying across observations i. It does not depend on the incidental parameters, α_i. Therefore, the conditional maximum-likelihood estimator of β 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 $\omega_i = 1$ given $y_{i1} + y_{i2} = 1$ is

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

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

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

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 β when 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

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

The information matrix is

$$J = E(J_{\tilde{B}_1}) = -\sum_{i=1}^{N} P_i F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \tag{7.3.18}$$
$$\{1 - F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]\}(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \cdot (\mathbf{x}_{i2} - \mathbf{x}_{i1})',$$

where $P_i = E(d_i \mid \alpha_i) = F(\beta'\mathbf{x}_{i1} + \alpha_i)[1 - F(\beta'\mathbf{x}_{i2} + \alpha_i)] + [1 - F(\beta'\mathbf{x}_{i1} + \alpha_i)]\ F(\beta'\mathbf{x}_{i2} + \alpha_i)$. 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,

$$\frac{1}{N} J_{\tilde{B}_1} - \frac{1}{N} J \text{ almost surely } \to 0 \text{ as } N \to \infty \tag{7.3.19}$$
$$\text{if } \sum_{i=1}^{N} \frac{1}{i^2} \mathbf{m}_i \mathbf{m}_i' < \infty,$$

where $\mathbf{m}_i$ replaces each element of $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ by 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 $D_{i1} = (d_{i11}, \ldots, d_{i1T}, \sum_{t=1}^{T} d_{i1t} = \sum_{t=1}^{T} y_{it} = s, 1 \le s \le T - 1)$, as the normalizing factor. Hence we may rewrite the conditional probability (7.3.14) as

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

Then the conditional log-likelihood function takes the form

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

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 β for the fixed-effects probit models.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

7.3.1.3 Some Monte Carlo Evidence

Given that there exists a consistent estimator of β for the fixed-effects logit model, but not for the fixed-effects probit model, and that in the binary case 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 β. 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

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

and

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

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 0 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 0. 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 β. The asymptotic properties of the fixed effects estimators of β 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 β, but it could be worse, namely,

[11] Similar results also hold for the MLE of the fixed-effects logit model. Wright and Douglas (1976), who used Monte Carlo methods to investigate the performance of the MLE, found that when $T = 20$, the MLE is virtually unbiased, and its distribution is well described by a limiting normal distribution, with the variance–covariance matrix based on the inverse of the estimated-information matrix.

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).

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 sampling from a univariate distribution G, indexed by a finite number of parameters δ, the log-likelihood function becomes

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

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 y conditional on α by a probability function that is marginal on α. It is a function of a finite number of parameters (β', δ'). Thus, maximizing (7.3.24), under weak regularity conditions, will give consistent estimators for β and δ as N tends to infinity provided the distribution (or conditional distribution) of α is correctly specified. If $G(\alpha)$ is misspecified, maximizing (7.3.24) will yield inconsistent estimates when T is fixed. However, when $T \to \infty$, the random effects estimator becomes consistent, irrespective of the form of the postulated distribution of individual effects. The reason is that:

$$\log f(\mathbf{y}_i \mid \mathbf{x}_i, \beta, \alpha_i) = \sum_{t=1}^{T} \log f(y_{it} \mid \mathbf{x}_{it}, \beta, \alpha_i)$$

is a sum of T time series observation, so that the distribution of α becomes negligible compared to that of the likelihood as the number of time periods increases (Arellano and Bonhomme 2009).

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 for α conditional on $\mathbf{x}$, $G(\alpha \mid \mathbf{x})$ and consider the marginal

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

log-likelihood function

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

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 $\mathbf{a}' = (\mathbf{a}_1', \ldots, \mathbf{a}_T)$ and $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, 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

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

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:

$$\begin{aligned} y_{it} &= 1 \text{ if } \beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta_i + u_{it} > 0, \quad t = 1, \ldots, T, \\ &= 0 \text{ if } \beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta_i + u_{it} \leq 0, \end{aligned} \tag{7.3.26}$$

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) or (7.3.25) are the same. So we shall discuss only the procedure to estimate the model (7.3.24).

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

mean α_i and variance 1, denoted by $\phi(v_{it} \mid \alpha_i)$ (Heckman 1981a). Then

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

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

where $\Phi(\cdot \mid \alpha_i)$ is the cumulative distribution function (cdf) of normal density with mean α_i and variance 1, $\phi(\cdot \mid \alpha_i)$, $c_{it} = -\beta' \mathbf{x}_{it}$, $b_{it} = \infty$ if $y_{it} = 1$ and $c_{it} = -\infty$, $b_{it} = -\beta' \mathbf{x}_{it}$ if $y_{it} = 0$, $G(\alpha_i \mid \mathbf{x}_i)$ 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, and 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 cumulative density functions for which highly accurate approximations are available. For instance, Butler and Moffitt (1982) suggests using Gaussian quadrature to achieve gains in computational efficiency. The Gaussian quadrature formula for evaluation of the necessary integral is the Hermite integration formula $\int_{-\infty}^{\infty} e^{-z^2} g(z) dz \approx \sum_{j=1}^{l} w_j g(z_j)$, 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 β, computationally it is much more 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:

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

Estimating each of t cross-sectional univariate probit specifications by maximum-likelihood gives the estimated coefficients of $\mathbf{x}_{it}$ and $\mathbf{x}_i$ as $\hat{\boldsymbol{\pi}}_t$, $t = 1, 2, \ldots, T$, which will converge to[12]

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

[12] In the case in which α_i are uncorrelated with $\mathbf{x}_i$, we have $\mathbf{a} = \mathbf{0}$ and $\sigma_\eta^2 = \sigma_\alpha^2$.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

as N tends to infinity where Π denotes the $T \times (T+1)K$ stacked $\boldsymbol{\pi}_t'$. Therefore, consistent estimators $(1+\sigma_\eta^2)^{-1/2}\beta'$ and $(1+\sigma_\eta^2)^{-1/2}\ \mathbf{a}'$ can be easily derived from (7.3.29). One can then follow Heckman's (1981a) suggestion by substituting these estimated values into (7.3.25) and optimizing the functions with respect to σ_η^2 conditional on $(1+\sigma_\eta^2)^{-1/2}\beta$ and $(1+\sigma_\eta^2)^{-1/2}\mathbf{a}$.

A more efficient estimator that avoids numerical integration is to impose the restriction (7.3.29) by $\boldsymbol{\pi} = \text{vec}\,(\Pi') = \mathbf{f}(\theta)$, where $\theta' = (\beta', \mathbf{a}', \sigma_\eta^2)$, and use a genralized method of moments (GMM) or minimum-distance estimator (see Chapter 3, Section 3.9), just as in the linear case. Chamberlain (1984) suggests that we choose $\hat{\theta}$ to minimize[13]

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

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

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

where

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

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

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

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

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

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 of estimating discrete choice model suffers from two drawbacks: (1) conditional on $\mathbf{x}$, the probability law of generating (u_{it}, α_i)

[13] Ω is the asymptotic variance–covariance matrix of $\hat{\boldsymbol{\pi}}$ when no restrictions are imposed on the variance–covariance matrix of the $T \times 1$ normal random variable $\mathbf{u}_i + \mathbf{e}\eta_i$. We can relax the serial-independence assumption on u_{it} and allow $E\mathbf{u}_i\mathbf{u}_i'$ to be an arbitrary positive definite matrix except for scale normalization. In this circumstance, $\Pi = \text{diag}\,\{(\sigma_{u1}^2 + \sigma_\eta^2)^{-1/2}, \dots, (\sigma_{uT}^2 + \sigma_\eta^2)^{-1/2}\}[I_T \otimes \beta' + \mathbf{e}\mathbf{a}']$.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

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 logit and linear probability model, there does not exist a simple transformation that can get rid of the incidental parameters. The semiparametric approach not only avoids making specific distribution of u_{it} but also allows consistent estimator of β 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

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

subject to the normalization condition $\mathbf{b}'\mathbf{b}$=1, where $\Delta\mathbf{x}_{it} = \mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, $\Delta y_{it} = y_{it} - y_{i,t-1}$, $sgn(w) = 1$ 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

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

where $H(\mathbf{b})$ is maximized at $\mathbf{b} = \beta^*$, where $\beta^* = \frac{\beta}{\|\beta\|}$ and $\| \beta \|$ is the square root of 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

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

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

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

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

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

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

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

$$H(\beta^*) - H(\mathbf{b}) = E\{[sgn(\Delta\mathbf{x}_{it}'\beta^*) - sgn(\Delta\mathbf{x}_{it}'\mathbf{b})](y_{it} - y_{i,t-1})\} \tag{7.4.6}$$

$$= 2\int_{W_b} sgn(\Delta\mathbf{x}_{it}'\beta^*)E[y_t - y_{t-1} \mid \Delta\mathbf{x}]dF_{\Delta\mathbf{x}},$$

where $W_b = [\Delta\mathbf{x} : sgn(\Delta\mathbf{x}'\beta^*) \neq sgn(\Delta\mathbf{x}'\mathbf{b})]$, 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 $\Delta\mathbf{x}$,

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

Therefore under the assumption that $\mathbf{x}$'s are unbounded,[14]

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

Manski (1985, 1987) has shown that under fairly general conditions, the estimator maximizing the criterion function (7.4.1) is a strongly consistent estimator for β^*.

As discussed in Chapter 3 and early sections of this chapter, when T is small the MLE of the (structural) parameters β 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 β in the nonlinear case. The Manski semiparametric approach makes 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 β.

The Manski maximum score estimator is consistent as $N \to \infty$ for unknown conditional distribution of u_{it} 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 shows that the maximum score estimator is probably not very useful in application because the properties of the limiting distribution are largely unknown.

The objective function (7.4.1) is equivalent to

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

subject to $\mathbf{b}'\mathbf{b} = 1$, $\mathbf{1}(A)$ 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 rate of 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

[14] If $\mathbf{x}$ is bounded, then identification may fail if u_{it} is not logistic (Chamberlain (2010)).

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

$H_N^*(\mathbf{b})$ 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) $\mid K(v) \mid < M$ 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 $\{h_N : N = 1, 2, \ldots\}$ be a sequence of strictly positive real numbers satisfying $\lim_{N \to \infty} h_N = 0$. Define

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

Horowitz (1992) defines a smoothed maximum score estimator as any solution that maximizes (7.4.9). Like Manski's estimator, β can be identified only up to scale. Instead of using the normalization $\| \beta^* \| = 1$, Horowitz (1992) finds it is more convenient to use the normalization that the coefficient of one component of $\Delta \mathbf{x}$, say $\Delta \mathbf{x}_1$, to be equal to 1 in absolute value if its coefficient $\beta_1 \neq 0$ and the probability distribution of $\Delta \mathbf{x}_1$ 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 0. The asymptotic behavior of the estimator can be analyzed 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 β 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 $\beta' \Delta \mathbf{x}$ 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 $h_N \to 0$. Lee (1999) suggests a root-N consistent semiparametric estimator that does not depend on a smoothing parameter by maximizing the double sums

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

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

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

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 $(\Delta y_{it} - \Delta y_{jt})\Delta y_{jt}^2 \Delta y_{it}^2 \neq 0$ excludes $(0, \pm 1)$ to make $\Delta y_{it} - \Delta y_{jt}$ binary (2 or -2). Conditional on given j, the first average over i and t converges to

$$E\{\, sgn(\Delta \mathbf{x}'\mathbf{b} - \Delta \mathbf{x}_j'\mathbf{b})(\Delta y - \Delta y_j)\Delta y^2 \Delta y_j^2 \mid \Delta \mathbf{x}_j, \Delta y_j\} \qquad (7.4.11)$$

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 $\sqrt{N}\left(\tilde{\mathbf{b}} - \tilde{\beta}\right)$ is equal to

$$4(E \bigtriangledown_2 \tau)^{-1}(E \bigtriangledown_1 \tau \bigtriangledown_1 \tau')(E \bigtriangledown_2 \tau)^{-1}, \qquad (7.4.12)$$

where $\tilde{\beta} = (\beta_2, \ldots, \beta_K)'$, and $\tilde{\mathbf{b}}$, its estimator,

$$\tau\left(\Delta y_j, \Delta \mathbf{x}_j, \tilde{\mathbf{b}}\right) \equiv E_{i|j}\{sgn(\Delta \mathbf{x}_i'\mathbf{b} - \Delta \mathbf{x}_j'\mathbf{b})(\Delta y_i - \Delta y_j)\Delta y_i^2 \Delta y_j^2\}, \quad i \neq j,$$

with $E_{i|j}$ denoting the conditional expectation of $(\Delta y_i, \Delta \mathbf{x}_i')$ conditional on $(\Delta y_j, \Delta \mathbf{x}_j')$, $\bigtriangledown_1 \tau$ and $\bigtriangledown_2 \tau$ denote the first and second derivative matrices of $\tau\left(\Delta y_j, \Delta \mathbf{x}_j, \tilde{\mathbf{b}}\right)$ with respect to $\tilde{\mathbf{b}}$.

The parametric approach requires the specification of the distribution of u. If the distribution of u is misspecified, the MLE of β is inconsistent. The semiparametric approach does not require the specification of the distribution of u and permits its distribution to depend on $\mathbf{x}$ in an unknown way (heteroskedasticity of unknown form). It is consistent up to a scale whether the unobserved individual effects are treated as fixed or correlated with $\mathbf{x}$. However, the step of differencing $\mathbf{x}_{it}$ eliminates time-invariant variables from the estimation. Lee's (1999) $\sqrt{N}$ consistent estimator takes the additional differencing across individuals, $\Delta \mathbf{x}_i - \Delta \mathbf{x}_j$, and further reduces the dimension of estimable parameters by eliminating "period individual-invariant" variables (e.g., time dummies and macroeconomic shocks common to all individuals) from the specification. Moreover, the requirement that u_{it} and $u_{i,t-1}$ are identically distributed conditional on $(\alpha_i, \mathbf{x}_{it}, \mathbf{x}_{i,t-1})$ does not allow the presence of the lagged dependent variables in $\mathbf{x}_{it}$. Neither can a semiparametric approach be used to generate the predicted probability conditional on $\mathbf{x}$ as in the parametric approach. All it can estimate is the relative effects of the explanatory variables.

7.5 DYNAMIC MODELS

7.5.1 The General Model

The static models discussed in the previous sections assume that the probability of moving (or staying) in or out of a state is independent of the occurrence or nonoccurrence of the event in the past. However, in a variety of contexts, such as in the study of the incidence of accidents (Bates and Neyman 1951), brand loyalty (Chintagunta, Kyriazidou, and Perktold 2001), labor force participation

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

(Heckman and Willis 1977; Hyslop 1999), and unemployment (Layton 1978), it is often noted that individuals who have experienced an event in the past are more likely to experience the event in the future than individuals who have not. In other words, the conditional probability that an individual will experience the event in the future is a function of past experience.

To analyze the intertemporal relationship among discrete variables, Heckman (1978a, 1981b) proposed a general framework in terms of a latent-continuous-random-variable crossing the threshold. He let the continuous random variable y_{it}^* be a function of $\mathbf{x}_{it}$ and past occurrence of the event,

$$y_{it}^* = \beta' \mathbf{x}_{it} + \sum_{l=1}^{t-1} \gamma_\ell y_{i,t-l} + \phi \sum_{s=1}^{t-1} \prod_{l=1}^{s} y_{i,t-l} + v_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T \tag{7.5.1}$$

and

$$y_{it} = \begin{cases} 1 & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \leq 0. \end{cases} \tag{7.5.2}$$

The error term v_{it} is assumed to be independent of $\mathbf{x}_{it}$ and is independently distributed over i, with a general intertemporal variance–covariance matrix $E\mathbf{v}_i\mathbf{v}_i' = \Omega$. The coefficient γ_l measures the effects of experience of the event l periods ago on current values of y_{it}^*. The coefficient ϕ measures the effect of the cumulative recent spell of experience in the state for those still in the state on the current value of y_{it}^*.

Specifications (7.5.1) and (7.5.2) accommodate a wide variety of stochastic models that appear in the literature. For example, let $\mathbf{x}_{it} = 1$, and let v_{it} be independently identically distributed. If $\gamma_l = 0, l = 2, \ldots, T-1$, and $\phi = 0$, equations (7.5.1) and (7.5.2) generate a time-homogenous first-order Markov process. If $\gamma_l = 0, 1, \ldots, T-1$, and $\phi \neq 0$, a renewal process is generated. If $\gamma_l = 0, l = 1, \ldots, T-1$ and $\phi = 0$, a simple Bernoulli model results. If one allows v_{it} to follow an autoregressive moving-average scheme, but keeps the assumption that $\gamma_l = 0, l = 1, \ldots, T-1$, and $\phi = 0$, the Coleman (1964) latent Markov model emerges.

As said before, repeated observations of a given group of individuals over time permit us to construct a model in which individuals may differ in their propensity to experience the event with the same $\mathbf{x}$. Such heterogeneity is allowed by decomposing the error term v_{it} as

$$v_{it} = \alpha_i + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{7.5.3}$$

where u_{it} is independently distributed over i, with arbitrary serial correlation, and α_i is individual-specific and can be treated as a fixed constant or as random. Thus, for example, if the previoius assumptions on the Markov process

$$\gamma_l = 0, l = 2, \ldots, T-1, \text{ and } \phi = 0$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

hold, but v_{it} follows a "components-of-variance" scheme (7.5.3), a compound first-order Markov process, closely related to previous work on the mover-stayer model (Goodman 1961; Singer and Spilerman 1976), is generated.

Specifications (7.5.1)–(7.5.3) allow for three sources of persistence (after controlling for the observed explanatory variables, $\mathbf{x}$). Persistence can be the result of serial correlation in the error term, u_{it}, or the result of "unobserved heterogeneity," α_i, or the result of true state dependence through the term $\gamma y_{i,t-1}$ or $\phi \prod_{l=1}^{*} y_{i,t-1}$. Distinguishing the sources of persistence is important because a policy that temporariliy increases the probability that $y = 1$ will have different implications about future probabilities of experiencing an event.

When the conditional probability of an individual staying in a state is a function of past experience, two new issues arise. One is how to treat the initial observations. The second is how to distinguish true state dependence from spurious state dependence in which the past y_{it} appears in the specification merely as a proxy for the unobserved individual effects, α_i. The first issue could play a role in deriving consistent estimators for a given model. The second issue is important because the time dependence among observed events could arise either from the fact that the actual experience of an event has modified individual behavior or from unobserved components that are correlated over time, or from a combination of both.

7.5.2 Initial Conditions

When dependence among time-ordered outcomes is considered, just as in the dynamic linear-regression model, the problem of initial conditions must be resolved for a likelihood approach before parameters generating the stochastic process can be estimated. To focus the discussion on the essential aspects of the problem of initial conditions and its solutions, we assume that there are no exogenous variables and that the observed data are generated by a first-order Markov process. Namely,

$$y_{it}^{*} = \beta_0 + \gamma y_{i,t-1} + v_{it}, \tag{7.5.4}$$

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

For ease of exposition we shall also assume that u_{it} is independently normally distributed with mean 0 and variance σ_u^2 normalized to be equal to 1. It should be noted that the general conclusions of the following discussion also hold for other types of distributions.

In much applied work in the social sciences, two assumptions for initial conditions are typically invoked: (1) the initial conditions or relevant presample history of the process are assumed to be truly exogenous, or (2) the process is assumed to be in equilibrium. Under the assumption that y_{i0} is a fixed nonstochastic constant for individual i, the joint probability of $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$,

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

given α_i, is

$$\prod_{t=1}^{T} F(y_{it} \mid y_{i,t-1}, \alpha_i) = \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\}, \quad (7.5.5)$$

where Φ is the standard normal cumulative distribution function. Under the assumption that the process is in equilibrium, the limiting marginal probability for $y_{it} = 1$ for all t, given α_i, is (Karlin and Taylor 1975)[15]

$$P_i = \frac{\Phi(\beta_0 + \alpha_i)}{1 - \Phi(\beta_0 + \gamma + \alpha_i) + \Phi(\beta_0 + \alpha_i)}, \quad (7.5.6)$$

and the limiting probability for $y_{it} = 0$ is $1 - P_i$. Thus the joint probability of $(y_{i0}, \ldots, y_{iT})$, given α_i is

$$\prod_{t=1}^{T} \Phi\{(\beta + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\} P_i^{y_{i0}} (1 - P_i)^{1 - y_{i0}}. \quad (7.5.7)$$

If α_i is random, with distribution $G(\alpha)$, the likelihood function for the random-effects model under the first assumption is

$$L = \prod_{i=1}^{N} \int \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\} dG(\alpha). \quad (7.5.8)$$

The likelihood function under the second assumption is

$$L = \prod_{i=1}^{N} \int \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\} \cdot P_i^{y_{i0}} (1 - P_i)^{1 - y_{i0}} dG(\alpha). \quad (7.5.9)$$

The likelihood functions (7.5.8) and (7.5.9) under both sets of assumptions about initial conditions are of closed form. When α_i is treated as random, the MLEs for β_0, γ, and σ_α^2 are consistent if N tends to infinity or if both N and

[15] The transition-probability matrix of our homogeneous two-state Markov chain is

$$\mathcal{P} = \begin{bmatrix} 1 - \Phi(\beta_0 + \alpha_i) & \Phi(\beta_0 + \alpha_i) \\ 1 - \Phi(\beta_0 + \gamma + a_i) & \Phi(\beta_0 + \gamma + \alpha_i) \end{bmatrix}.$$

By mathematical induction, the n-step transition matrix is

$$\begin{aligned} \mathcal{P}^n = {} & \frac{1}{1 - \Phi(\beta_0 + \gamma + \alpha_i) + \Phi(\beta_0 + \alpha_i)} \\ & \cdot \left\{ \begin{bmatrix} 1 - \Phi(\beta_0 + \gamma + \alpha_i) & \Phi(\beta_0 + \alpha_i) \\ 1 - \Phi(\beta_0 + \gamma + \alpha_i) & \Phi(\beta_0 + \alpha_i) \end{bmatrix} \right. \\ & + [\Phi(\beta_0 + \gamma + \alpha_i) - \Phi(\beta_0 + \alpha))]^n \\ & \left. \cdot \begin{bmatrix} \Phi(\beta_0 + \alpha_i) & -\Phi(\beta_0 + \alpha_i) \\ -[1 - \Phi(\beta_0 + \gamma + \alpha_i)] & 1 - \Phi(\beta_0 + \gamma + \alpha_i) \end{bmatrix} \right\}. \end{aligned}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

T tend to infinity. When α_i is treated as a fixed constant (7.5.5), the MLEs for β_0, γ, and α_i are consistent only when T tends to infinity. If T is finite, the MLE is biased. Moreover, the limited results from Monte Carlo experiments suggest that, contrary to the static case, the bias is significant (Heckman 1981b).

However, the assumption that initial conditions are fixed constants is justifiable only if the disturbances that generate the process are serially independent and if a genuinely new process is fortuitously observed at the beginning of the sample. If the process has been in operation prior to the time it is sampled, or if the disturbances of the model are serially dependent as in the presence of individual specific random effects, the initial conditions are not exogenous. The assumption that the process is in equilibrium also raises problems in many applications, especially when time-varying exogenous variables are driving the stochastic process.

Suppose that the analyst does not have access to the process from the beginning; then the initial state for individual i, y_{i0}, cannot be assumed fixed. The initial state is determined by the process generating the panel sample. The sample likelihood function for the fixed-effects model is

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\}\ f\ (y_{i0} \mid \alpha_i), \tag{7.5.10}$$

and the sample likelihood function for the random-effects models is

$$L = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\}\ f\ (y_{i0} \mid \alpha) dG(\alpha), \tag{7.5.11}$$

where $f(y_{i0} \mid \alpha)$ denotes the marginal probability of y_{i0} given α_i. Thus, unless T is very large, maximizing (7.5.5) or (7.5.10) yields inconsistent estimates.[16]

Because y_{i0} is a function of unobserved past values, besides the fact that the marginal distribution of $f(y_{i0} \mid \alpha)$ is not easy to derive, maximizing (7.5.10) or (7.5.11) is also considerably involved. Heckman (1981b) therefore suggested that we approximate the initial conditions for a dynamic discrete model by the following procedure:

1. Approximate the probability of y_{i0}, the initial state in the sample, by a probit model, with index function

$$y_{i0}^* = Q(\mathbf{x}_{it}) + \epsilon_{i0}, \tag{7.5.12}$$

and

$$y_{i0} = \begin{cases} 1 & \text{if } y_{i0}^* > 0, \\ 0 & \text{if } y_{i0}^* \leq 0, \end{cases} \tag{7.5.13}$$

[16] This can be easily seen by noting that the expectation of the first-derivative vector of (7.5.5) or (7.5.8) with respect to the structural parameters does not vanish at the true parameter value when the expectations are evaluated under (7.5.10) or (7.5.11).

where $Q(\mathbf{x}_{it})$ is a general function of $\mathbf{x}_{it}$, $t = 0, \ldots, T$, usually specified as linear in $\mathbf{x}_{it}$, and ϵ_{i0} is assumed to be normally distributed, with mean 0 and variance 1.

2. Permit ϵ_{i0} to be freely correlated with v_{it}, $t = 1, \ldots, T$.
3. Estimate the model by maximum likelihood without imposing any restrictions between the parameters of the structural system and parameters of the approximate reduced-form probability for the initial state of the sample.

Heckman (1981b) conducted Monte Carlo studies comparing the performances of the MLEs when assumption on initial y_{i0} and α_i conform with the true data generating process, an approximate reduced-form probability for y_{i0}, and false fixed y_{i0} and α_i for a first-order Markov process. The data for his experiment were generated by the random-effects model

$$y_{it}^* = \beta x_{it} + \gamma y_{i,t-1} + \alpha_i + u_{it}, \tag{7.5.14}$$

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

where the exogenous variable x_{it} was generated by (7.3.23). He let the process operate for 25 periods before selecting samples of 8 ($= T$) periods for each of the 100 ($=N$) individuals used in the 25 samples for each parameter set. Heckman's Monte Carlo results are produced in Table 7.2.

These results show that contrary to the static model, the fixed-effects probit estimator performs poorly. The greater the variance of the individual effects (σ_α^2), the greater the bias. The t statistics based on the estimated information matrix also lead to a misleading inference by not rejecting the false null hypotheses of $\gamma = \beta = 0$ in the vast majority of samples.

By comparison, Heckman's approximate solution performs better. Although the estimates are still biased from the true values, their biases are not significant, particularly when they are compared with the ideal estimates. The t statistics based on the approximate solutions are also much more reliable than in the fixed-effects probit model, because they lead to a correct inference in a greater proportion of the samples.

Heckman's Monte Carlo results also point to a disquieting feature. Namely, the MLE produces a biased estimator even under the ideal conditions with a correctly specified likelihood function. Because a panel with 100 observations of three periods is not uncommon, this finding deserves further study.

7.5.3 A Conditional Approach

The likelihood approach cannot yield a consistent estimator when T is fixed and N tends to infinity if the individual effects are fixed. If the individual effects are random and independent of $\mathbf{x}$, the consistency of the MLE depends on the correct formulation of the probability distributions of the effects and initial observations. A semiparametric approach cannot be implemented for

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Table 7.2. *Monte Carlo results for first-order Markov process*

		$\sigma_\alpha^2 = 3$			$\sigma_\alpha^2 = 1$		
γ		$\beta = -0.1$	$\beta = 1$	$\beta = 0$	$\beta = -0.1$	$\beta = 1$	$\beta = 0$
Values of $\hat{\gamma}$ and $\hat{\beta}$ for the random-effects estimator with known initial conditions[a]							
0.5	$\hat{\gamma}$	n.a.[c]	0.57	n.a.[c]			
	$\hat{\beta}$	n.a.[c]	0.94	—[d]			
0.1	$\hat{\gamma}$	0.13	0.12	0.14			
	$\hat{\beta}$	−0.11	1.10	—			
Values of $\hat{\gamma}$ and $\hat{\beta}$ for the approximate random-effects estimation[a]							
0.5	$\hat{\gamma}$	0.63	0.60	0.70	n.a.[c]	0.54	0.62
	$\hat{\beta}$	−0.131	0.91	—	n.a.[c]	0.93	—
0.1	$\hat{\gamma}$	0.14	0.13	0.17	0.11	0.11	0.13
	$\hat{\beta}$	−0.12	0.92	—	−0.12	0.95	—
Values of $\hat{\gamma}$ and $\hat{\beta}$ for the fixed-effects estimator[b]							
0.5	$\hat{\gamma}$	0.14	0.19	0.03	n.a.[c]	0.27	0.17
	$\hat{\beta}$	−0.07	1.21	—	n.a.[c]	1.17	—
0.1	$\hat{\gamma}$	−0.34	−0.21	−0.04	−0.28	−0.15	−0.01
	$\hat{\beta}$	−0.06	1.14	—	−0.08	1.12	—

[a] $N = 100; T = 3$.
[b] $N = 100; T = 8$.
[c] Data not available because the model was not estimated.
[d] Not estimated.
Source: Heckman (1981b, Table 4.2).

a dynamic model because the strict exogeneity condition of explanatory variables is violated with the presence of lagged dependent variables as explanatory variables. When strict exogenity condition of the explanatory variables is violated, $E(\Delta u_{it} \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, y_{i,t-1}, y_{i,t-2}) \neq 0$. In other words, the one-to-one correspondence relation of the form (7.4.4) is violated. Hence, the Manski (1985) type maximum score estimator cannot be implemented. Neither can the (unrestricted) conditional approach be implemented. Consider the case of $T = 2$. The basic idea of conditional approach is to consider the probability of $y_{i2} = 1$ or 0 conditional on explanatory variables in both periods and conditional on $y_{i1} \neq y_{i2}$. If the explanatory variables of Prob $(y_{i2} = 1)$ include y_{i1}, then the conditional probability is either 1 or 0 according as $y_{i1} = 0$ or 1, hence provides no information about γ and β.

However, in the case that $T \geq 3$ and $\mathbf{x}_{it}$ follows certain special pattern. Honoré and Kyriazidou (2000a) show that it is possible to generalize the conditional probability approach to consistently estimate the unknown parameters for the logit model or to generalize the maximum score approach without the

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

need of formulating the distribution of α_i or the probability distribution of the initial observations for certain types of discrete choice models. However, the estimators converge to the true values at the speed considerably slower than the usual square root N rate.

Consider the model (7.5.4) with the assumption that u_{it} is logistically distributed, then the model of $(y_{i0}, \ldots, y_{iT})$ is of the form

$$P(y_{i0} = 1 \mid \alpha_i) = P_0(\alpha_i) \tag{7.5.15}$$

$$P(y_{it} = 1 \mid \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp(\gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\gamma y_{i,t-1} + \alpha_i)}, \tag{7.5.16}$$

for $t = 1, 2, \ldots, T$.

When $T \geq 3$, Chamberlain (1993) has shown that inference on γ can be made independent of α_i by using a conditional approach.

For ease of exposition, we shall assume that $T = 3$ (i.e., there are four time series observations for each i). Consider the events

$$\begin{aligned} A &= \{y_{i0}, y_{i1} = 0, y_{i2} = 1, y_{i3}\}, \\ B &= \{y_{i0}, y_{i1} = 1, y_{i2} = 0, y_{i3}\}. \end{aligned}$$

where y_{i0} and y_{i3} can be either 1 or 0. Then

$$\begin{aligned} P(A) = P_0(\alpha_i)^{y_{i0}}[1 - P_0(\alpha_i)]^{1-y_{i0}} \cdot \frac{1}{1 + \exp(\gamma y_{i0} + \alpha_i)} \\ \cdot \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \cdot \frac{\exp(y_{i3}(\gamma + \alpha_i))}{1 + \exp(\gamma + \alpha_i)} \end{aligned} \tag{7.5.17}$$

and

$$\begin{aligned} P(B) = P_0(\alpha_i)^{y_{i0}}[1 - P_0(\alpha_i)]^{1-y_{i0}} \cdot \frac{\exp(\gamma y_{i0} + \alpha_i)}{1 + \exp(\gamma y_{i0} + \alpha_i)} \\ \cdot \frac{1}{1 + \exp(\gamma + \alpha_i)} \cdot \frac{\exp(\alpha_i y_{i3})}{1 + \exp(\alpha_i)}. \end{aligned} \tag{7.5.18}$$

Hence

$$\begin{aligned} P(A \mid A \cup B) &= P(A \mid y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) \\ &= \frac{\exp(\gamma y_{i3})}{\exp(\gamma y_{i3}) + \exp(\gamma y_{i0})} \\ &= \frac{1}{1 + \exp[\gamma(y_{i0} - y_{i3})]}, \end{aligned} \tag{7.5.19}$$

and

$$\begin{aligned} P(B \mid A \cup B) &= P(B \mid y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) \\ &= 1 - P(A \mid A \cup B) \\ &= \frac{\exp[\gamma(y_{i0} - y_{i3})]}{1 + \exp[\gamma(y_{i0} - y_{i3})]}. \end{aligned} \tag{7.5.20}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Equation (7.5.19) and (7.5.20) are in the binary logit form and does not depend on α_i. The conditional log-likelihood

$$\log \tilde{L} = \sum_{i=1}^{N} 1(y_{i1} + y_{i2} = 1)\{y_{i1}[\gamma(y_{i0} - y_{i3})] - \log [1 + \exp \gamma(y_{i0} - y_{i3})]\} \tag{7.5.21}$$

is in the conditional logit form. Maximizing (7.5.21) yields $\sqrt{N}$ consistent estimator of γ, where $1(A) = 1$ if A occurs and 0 otherwise.

When exogenous variables $\mathbf{x}_{it}$ also appear as explanatory variables in the latent response function

$$y_{it}^{*} = \beta' \mathbf{x}_{it} + \gamma y_{i,t-1} + \alpha_i + u_{it}, \tag{7.5.22}$$

we may write

$$P(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = P_0(\mathbf{x}_i, \alpha_i), \tag{7.5.23}$$

$$\begin{aligned} P(y_{it} &= 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) \\ &= \frac{\exp (\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)}{1 + \exp (\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)}, \quad t = 1, \ldots, T. \end{aligned} \tag{7.5.24}$$

Let $P(y_{i0}) = P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}}$. Suppose $T = 3$. Under (7.5.24),

$$\begin{aligned} P(A) = P(y_{i0}) \cdot &\frac{1}{1 + \exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)} \cdot \frac{\exp (\mathbf{x}_{i2}'\boldsymbol{\beta} + \alpha_i)}{1 + \exp (\mathbf{x}_{i2}'\beta + \alpha_i)} \\ &\cdot \frac{\exp [(\mathbf{x}_{i3}'\boldsymbol{\beta} + \gamma + \alpha_i) y_{i3}]}{1 + \exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \gamma + \alpha_i)}. \end{aligned} \tag{7.5.25}$$

$$\begin{aligned} P(B) = P(y_{i0}) \cdot &\frac{\exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)}{1 + \exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)} \\ &\cdot \frac{1}{1 + \exp (\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma + \alpha_i)} \cdot \frac{[\exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \alpha_i) y_{i3}]}{1 + \exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \alpha_i)}. \end{aligned} \tag{7.5.26}$$

The denominator of $P(A)$ and $P(B)$ are different depending the sequence is of $(y_{i1} = 0, y_{i2} = 1)$ or $(y_{i1} = 1, y_{i2} = 0)$. Therefore, in general, $P(A \mid \mathbf{x}_i, \alpha_i, A \cup B)$ will depend on α_i. However, if $\mathbf{x}_{i2} = \mathbf{x}_{i3}$, then the denominator of $P(A)$ and $P(B)$ are identical. Using the same conditioning method, Honoré and Kyriazidou (2000a) show that

$$\begin{aligned} P(A \mid \mathbf{x}_i, \alpha_i, A \cup B, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ = \frac{1}{1 + \exp [(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta + \gamma(y_{i0} - y_{i3})]}, \end{aligned} \tag{7.5.27}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

which does not depend on α_i. If $\mathbf{x}_{it}$ is continuous, it may be rare that $\mathbf{x}_{i2} = \mathbf{x}_{i3}$. Honoré and Kyriazidou (2000a) propose estimating β and γ by maximizing

$$\sum_{i=1}^{N} \mathbf{1}(y_{i1} + y_{i2} = 1) K\left(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N}\right) \ln\left\{\frac{\exp\left[(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i0} - y_{i3})\right]^{y_{i1}}}{1 + \exp\left[(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i0} - y_{i3})\right]}\right\} \tag{7.5.28}$$

with respect to $\boldsymbol{\beta}$ and γ (over some compact set) if $P(\mathbf{x}_{i2} = \mathbf{x}_{i3}) > 0$. Here $K(\cdot)$ is a kernel density function that gives appropriate weight to observation i, while h_N is a bandwidth which shrinks to 0 as N tends to infinity. The asymptotic theory will require that $K(\cdot)$ be chosen so that a number of regularity conditions are satisfied such as $\mid K(\cdot) \mid < M$ for some constant M, and $K(v) \to 0$ as $\mid v \mid \to \infty$ and $\int K(v)dv = 1$. For instance, $K(v)$ is often taken to be the standard normal density function and $h_N = cN^{-1/5}$ for some constant c. The effect of the term $K(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N})$ is to give more weight to observations for which $\mathbf{x}_{i2}$ is close to $\mathbf{x}_{i3}$. Their estimator is consistent and asymptotically normal although their speed of convergence is only $\sqrt{Nh_N^k}$, which is considerably slower than $\sqrt{N}$ where k is the dimension of $\mathbf{x}_{it}$.

The conditional approach works for the logit model but it does not seem applicable for general nonlinear models. However, if the nonlinearity can be put in the single index form $F(a)$ with the transformation function F being a strictly increasing distribution function, then Manski (1987) maximum score estimator for the static case can be generalized to the case where the lagged dependent variable is included in the explanatory variable set by considering

$$\begin{aligned} &P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ &= P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}} \\ &\quad \cdot [1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)] \times F(\mathbf{x}_{i2}'\beta + \alpha_i) \\ &\quad \cdot [1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]^{1-y_{i3}} \times F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)^{y_{i3}} \end{aligned} \tag{7.5.29}$$

and

$$\begin{aligned} &P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ &= P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}} \\ &\quad \cdot F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i) \times [1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)] \\ &\quad \cdot [1 - F(\mathbf{x}_{i2}'\beta + \alpha_i)]^{1-y_{i3}} \times F(\mathbf{x}_{i2}'\beta + \alpha_i)^{y_{i3}}. \end{aligned} \tag{7.5.30}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

If $y_{i3} = 0$, then

$$\begin{aligned}
&\frac{P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})}{P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)},
\end{aligned} \tag{7.5.31}$$

where the second equality follows from the fact that $y_{i3} = 0$. If $y_{i3} = 1$, then

$$\begin{aligned}
&\frac{P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})}{P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i))]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)},
\end{aligned} \tag{7.5.32}$$

where the second equality follows from the fact that $y_{i3} = 1$, so that $\gamma y_{i3} = \gamma$. In either case, the monotoncity of F implies that

$$\frac{P(A)}{P(B)} \begin{cases} > 1 \text{ if } \mathbf{x}_{i2}'\beta + \gamma y_{i3} > \mathbf{x}_{i1}'\beta + \gamma y_{i0}, \\ < 1 \text{ if } \mathbf{x}_{i2}'\beta + \gamma y_{i3} < \mathbf{x}_{i1}\beta + \gamma y_{i0}. \end{cases}$$

Therefore,

$$\begin{aligned}
&\operatorname{sgn}[P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) - P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})] \\
&\quad = \operatorname{sgn}[(\mathbf{x}_{i2} - \mathbf{x}_{i1})'\beta + \gamma(y_{i3} - y_{i0})].
\end{aligned} \tag{7.5.33}$$

Hence, Honoré and Kyriazidou (2000a) propose a maximum score estimator that maximizes the score function

$$\sum_{i=1}^{N} K\left(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N}\right)(y_{i2} - y_{i1})\operatorname{sgn}[(\mathbf{x}_{i2} - \mathbf{x}_{i1})'\beta + \gamma(y_{i3} - y_{i0})] \tag{7.5.34}$$

with respect to β and γ. The Honoré and Kyriazidou estimator is consistent (up to a scale) if the density of $\mathbf{x}_{i2} - \mathbf{x}_{i3}$, $f(\mathbf{x}_{i2} - \mathbf{x}_{i3})$, is strictly positive at 0, $f(0) > 0$. (This assumption is required for consistency.)

We have discussed the estimation of panel data dynamic discrete choice model assuming that $T = 3$. It can be easily generalized to the case of $T > 3$ by maximizing the objective function that is based on sequences where an individual switches between alternatives in any two of the middle $T - 1$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

periods:

$$\sum_{i=1}^{N} \sum_{1 \le s < t \le T-1} \mathbf{1}\{y_{is} + y_{it} = 1\} K\left(\frac{\mathbf{x}_{i,t+1} - \mathbf{x}_{i,s+1}}{h_N}\right)$$
$$\cdot \ln\left(\frac{\exp\left[(\mathbf{x}_{is} - \mathbf{x}_{it})'\beta + \gamma(y_{i,s-1} - y_{i,t+1}) + \gamma(y_{i,s+1} - y_{i,t-1})\mathbf{1}(t-s>\mathbf{1})\right]^{y_{is}}}{\mathbf{1} + \exp\left[(\mathbf{x}_{is} - \mathbf{x}_{it})'\beta + \gamma(y_{i,s-1} - y_{i,t+1}) + \gamma(y_{i,s+1} - y_{i,t-1})\mathbf{1}(t-s>\mathbf{1})\right]}\right) \tag{7.5.35}$$

The conditional approach does not require modeling of the initial observations of the sample. Neither does it make any assumptions about the statistical relationship of the individual effects with the observed explanatory variables or with the initial conditions. However, it also suffers from the limitation that $\mathbf{x}_{is} - \mathbf{x}_{it}$ has support in a neighborhood of 0 for any $t \neq s$, which rules out time dummies as explanatory variables.[17] The fact that individual effects cannot be estimated also means that it is not possible to carry out predictions or compute elasticities for individual agents at specified values of the explanatory variables.

7.5.4 State Dependence versus Heterogeneity

There are two diametrically opposite explanations for the often observed empirical regularity with which individuals who have experienced an event in the past are more likely to experience that event in the future. One explanation is that as a consequence of experiencing an event, preferences, prices, or constraints relevant to future choices are altered. A second explanation is that individuals may differ in certain unmeasured variables that influence their probability of experiencing the event but are not influenced by the experience of the event. If these variables are correlated over time and are not properly controlled, previous experience may appear to be a determinant of future experience solely because it is a proxy for such temporally persistent unobservables. Heckman (1978a, 1981a,c) has termed the former case "true state dependence" and the latter case "spurious state dependence," because in the former case, past experience has a genuine behavioral effect in the sense that an otherwise identical individual who has not experienced the event will behave differently in the future than an individual who has experienced the event. In the latter case, previous experience appears to be a determinant of future experience solely because it is a proxy for temporally persistent unobservables that determine choices.

The problem of distinguishing between true and spurious state dependencies is of considerable substantive interest. To demonstrate this, let us consider some work in the theory of unemployment. Phelps (1972) argued that current

[17] See Arellano and Carrasco (2003) for a GMM approach to estimate the dynamic random-effects probit model.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

unemployment has a real and lasting effect on the probability of future unemployment. Hence, short-term economic policies that alleviate unemployment tend to lower aggregate unemployment rates in the long run by preventing the loss of work-enhancing market experience. On the other hand, Cripps and Tarling (1974) maintained the opposite view in their analysis of the incidence and duration of unemployment. They assumed that individuals differ in their propensity to experience unemployment and in their unemployment duration times and those differences cannot be fully accounted for by measured variables. They further assumed that the actual experience of having been unemployed or the duration of past unemployment does not affect future incident or duration. Hence, in their model, short-term economic policies have no effect on long-term unemployment.

Because the unobserved individual effects, α_i, persist over time, ignoring these effects of unmeasured variables (heterogeneity) creates serially correlated residuals. This suggests that we cannot use the conditional probability, given past occurrence not equal to the marginal probability alone, Prob $(y_{it} \mid y_{i,t-s}, \mathbf{x}_{it}) \neq$ Prob $(y_{it} \mid \mathbf{x}_{it})$, to test for true state dependence against spurious state dependence, because this inequality may be a result of past information on y yielding information on the unobserved specific effects. A proper test for dependence should control for the unobserved individual-specific effects.

When conditional on the individual effects, α_i, the error term u_{it} is serially uncorrelated, a test for state dependence can be implemented by controlling the individual effects and testing for the conditional probability equal to the marginal probability,

$$\text{Prob}\,(y_{it} \mid y_{i,t-s}, \mathbf{x}_{it}, \alpha_i) = \text{Prob}(y_{it} \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.36}$$

When N is fixed and $T \to \infty$, likelihood ratio tests can be implemented to test (7.5.36).[18] However, if T is finite, controlling α_i to obtain consistent estimator for the coefficient of lagged dependent variable imposes very restrictive conditions on the data which severly limits the power of the test, as shown in Section 7.4.

If α_i are treated as random and the conditional distribution of α_i given $\mathbf{x}_i$ is known, a more powerful test is to use an unconditional approach. Thus, one may test true state dependence versus spurious state dependence by testing the

[18] Let $P_{it} = \text{Prob}\,(y_{it} \mid \mathbf{x}_{it}, \alpha_i)$ and $P^*_{it} = \text{Prob}\,(y_{it} \mid y_{i,t-\ell}, \mathbf{x}_{it}, \alpha_i)$. Let $\hat{P}_{it}$ and $\hat{P}^*_{it}$ be the MLEs obtained by maximizing $\mathcal{L} = \prod_i \prod_t P_{it}^{y_{it}}(1 - P_{it})^{1-y_{it}}$ and $\mathcal{L}^* = \prod_i \prod_t P_{it}^{*y_{it}}(1 - P^*_{it})^{1-y_{it}}$ with respect to unknown parameters, respectively. A likelihood-ratio test statistic for the null hypothesis (7.5.36) is $-2 \log [\mathcal{L}(\hat{P}_{it})/\mathcal{L}(\hat{P}^*_{it})]$. When conditional on $\mathbf{x}_{it}$ and α_i, there are repeated observations; we can also use the Pesaran chi-square goodness-of-fit statistic to test (7.5.36). For details, see Bishop, Fienberg, and Holland (1975, Chapter 7). However, in the finite-T case, the testing procedure cannot be implemented, as the $\alpha_i's$ are unknown and cannot be consistently estimated.

significance of the MLE of γ of the log-likelihood

$$\sum_{i=1}^{N} \log \int \prod_{t=1}^{T} \left\{ F(\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)^{y_{it}} \left[1 - F(\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i) \right]^{1-y_{it}} \right. \tag{7.5.37}$$

$$\left. \cdot P(\mathbf{x}_i, \alpha)^{y_{i0}} \left[1 - P(\mathbf{x}_i, \alpha) \right]^{1-y_{i0}} \right\} G(\alpha_i \mid \mathbf{x}_i) d\alpha_i .$$

When conditional on the individual effects, α_i, the error term u_{it} remains serially correlated, the problem becomes more complicated. The conditional probability, Prob $(y_{it} \mid y_{i,t-1}, \alpha_i)$, not being equal to the marginal probability, Prob $(y_{it} \mid \alpha_i)$, could be because of past y_{it} containing information on u_{it}. A test for state dependence cannot simply rely on the multinominal distribution of the $(y_{i1}, \ldots, y_{iT})$ sequence. The general framework (7.5.1) and (7.5.2) proposed by Heckman (1978a, 1981a,b) accommodates very general sorts of heterogeneity and structural dependence. It permits an analyst to combine models and test among competing specifications within a unified framework. However, the computations of maximum-likelihood methods for the general models could be quite involved. It would be useful to rely on simple methods to explore data before implementing the computationally cumbersome maximum-likelihood method for a specific model.

Chamberlain (1978b) suggested a simple method to distinguish true state dependence from spurious state dependence. He noted that just as in the continuous models, a key distinction between state dependence and serial correlation is whether or not there is a dynamic response to an intervention. This distinction can be made clear by examining (7.5.1). If $\gamma = 0$, a change in $\mathbf{x}$ has its full effect immediately, whereas if $\gamma \neq 0$, this implies a distributed-lag response to a change in $\mathbf{x}$. The lag structure relating y to $\mathbf{x}$ is not related to the serial correlation in u. If $\mathbf{x}$ is increased in period t and then returned to its former level, the probability of $y_{i,t+1}$ is not affected if $\gamma = 0$, because by assumption the distribution of u_{it} was not affected. If $\gamma \neq 0$, then the one-period shift in $\mathbf{x}$ will have lasting effects. An intervention that affects the probability of y in period t will continue to affect the probability of y in period $t + 1$, even though the intervention was presented only in period t. In contrast, an interpretation of serial correlation is that the shocks (u) tend to persist for more than one period and that $y_{i,t-s}$ is informative only in helping to infer u_{it} and hence to predict u_{it}. Therefore, a test that is not very sensitive to functional form is to simply include lagged $\mathbf{x}'s$ without lagged y. After conditioning on the individual-specific effect α, there may be two possible outcomes. If there is no state dependence, then

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots, \alpha_i) = \text{ Prob } (y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.38}$$

If there is state dependence, then

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots, \alpha_i) \neq \text{ Prob } (y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.39}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

While the combination of (7.5.38) and (7.5.39) provides a simple form to distinguish pure heterogeneity, state dependence, and serial correlation, we cannot make further distinctions with regard to different forms of state dependence and heterogeneity, and serial correlation should (7.5.38) be rejected. Models of the form (7.5.1) and (7.5.2) may still have to be used to further narrow down possible specifications.

7.5.5 Two Examples

The control of heterogeneity plays a crucial role in distinguishing true state dependence from spurious state dependence. Neglecting heterogeneity and the issue of initial observations can also seriously bias the coefficient estimates. It is important in estimating dynamic models that the heterogeneity in the sample be treated correctly. To demonstrate this, we use the female-employment models estimated by Heckman (1981c) and household brand choices estimated by Chintagunta, Kyriazidou, and Perktold (2001) as examples.

7.5.5.1 Female Employment

Heckman (1981c) used the first three-year sample of women aged 45–59 in 1968 from the Michigan Panel Survey of Income dynamics to study married women's employment decisions. A woman is defined to be a market participant if she worked for money any time in the sample year. The set of explanatory variables is as follows: the woman's education; family income, excluding the wife's earnings; number of children younger than six; number of children at home; unemployment rate in the county in which the woman resided; the wage of unskilled labor in the county (a measure of the availability of substitutes for a woman's time in the home); the national unemployment rate for prime-age men (a measure of aggregate labor-market tightness); two types of prior work experience: within-sample work experience and presample work experience. The effect of previous work experience is broken into two components, because it is likely that presample experience exerts a weaker measured effect on current participation decisions than more recent experience. Furthermore, because the data on presample work experience are based on a retrospective question and therefore are likely to be measured with error, Heckman replaces them by predicted values based on a set of regressors.

Heckman fitted the data to various multivariate probit models of the form (7.5.1) and (7.5.2) to investigate whether or not work experience raises the probability that a woman will work in the future (by raising her wage rates) and to investigate the importance of controlling for heterogeneity in utilizing panel data. Maximum-likelihood-coefficient estimates for the state-dependent models under the assumptions of stationary intertemporal covariance matrix

$$\Omega = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ & 1 & \rho_{23} \\ & & 1 \end{bmatrix},$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

first-order Markov process ($v_{it} = \rho v_{i,t-1} + u_{it}$), and no heterogeneity ($v_{it} = u_{it}$) are presented in columns 1, 2, and 3, respectively, of Table 7.3.[19] Coefficient estimates for no state dependence with general stationary intertemporal correlation, first-order Markov process, conventional random effects error-component formulation $v_{it} = \alpha_i + u_{it}$, equivalent to imposing the restriction that $\rho_{12} = \rho_{13} = \rho_{23} = \sigma_\alpha^2/(\sigma_u^2 + \sigma_\alpha^2)$, and no heterogeneity are presented in columns 4, 5, 6, and 7, respectively. A Heckman–Willis (1977) model with time-invariant exogenous variables and conventional error-component formulation was also estimated and is presented in column 8.

Likelihood ratio test statistics (twice the difference of the log-likelihood value) against the most general model (column 1, Table 7.3) indicate the acceptance of recent labor-market experience as an important determinant of current employment decision, with unobservables determining employment choices following a first-order Markov process (column 2, Table 7.3) as a maintained hypothesis, and the statistics clearly reject all other formulations. In other words, Heckman's study found that work experience, as a form of general and specific human capital investment, raises the probability that a woman will work in the future, even after accounting for serial correlation of a very general type. It also maintained that there exist unobserved variables that affect labor participations. However, initial differences in unobserved variables tend to be eliminated with the passage of time. But this homogenizing effect is offset in part by the impact of prior work experience that tends to accentuate initial differences in the propensity to work.

Comparison of the estimates of the maintained hypothesis with estimates of other models indicates that the effect of recent market experience on employment is dramatically overstated in a model that neglects heterogeneity. The estimated effect of recent market experience on current employment status recorded in column 3, Table 7.3, overstates the impact by a factor of 10 (1.46 vs. 0.143)! Too much credit will be attributed to past experience as a determinant of employment if intertemporal correlation in the unobservables is ignored. Likewise for the estimated impact of national unemployment on employment. On the other hand, the effect of children on employment is understated in models that ignore heterogeneity.

Comparisons of various models' predictive performance on sample-run patterns (temporal employment status) are presented in Table 7.4. It shows that dynamic models ignoring heterogeneity under-predict the number of individuals who work all of the time and over-predict the number who do not work at all. It also overstates the estimated frequency of turnover in the labor force. In fact, comparing the performances of the predicted run patterns for the dynamic and static models without heterogeneity (column 3 and 7 or Table 7.3 and columns 3 and 4 of Table 7.4) suggests that introducing "lagged employment status" into a model as a substitute for a more careful treatment of heterogeneity is an imperfect procedure. In this case, it is worse than using no proxy at all. Nor

[19] A nonstationary model was also estimated by Heckman (1981c), but because the data did not reject stationarity, we shall treat the model as having stationary covariance.

Table 7.3. *Estimates of employment models for women aged 45–59 in 1968*[a]

Variable	(1)	(2)	(3)
Intercept	−2.576 (4.6)	1.653 (2.5)	0.227 (0.4)
No. of children aged<6	−0.816 (2.7)	−0.840 (2.3)	−0.814 (2.1)
County unemployment rate (%)	−0.035 (1.5)	−0.027 (1.0)	−0.018 (0.57)
County wage rate ($/h)	0.104 (0.91)	0.104 (0.91)	0.004 (0.02)
Total no. of children	−0.146 (4.3)	−0.117 (2.2)	−0.090 (2.4)
Wife's education (years)	0.162 (6.5)	0.105 (2.8)	0.104 (3.7)
Family income, excluding wife's earnings	-0.363×10^{-4} (4.8)	-0.267×10^{-4} (2.7)	-0.32×10^{-4} (3.6)
National unemployment rate	−0.106 (0.51)	−0.254 (1.4)	−1.30 (6)
Recent experience	0.143 (0.95)	0.273 (1.5)	1.46 (12.2)
Predicted presample experience	0.072 (5.8)	0.059 (3.4)	0.045 (3.4)
Serial-correlation coefficient:			
ρ_{12}	0.913	—	—
ρ_{13}	0.845		
ρ_{23}	0.910		
ρ	—	0.873 (14.0)	—
$\sigma_\alpha^2/(\sigma_u^2+\sigma_\alpha^2)$	—	—	—
Log likelihood	−237.74	−240.32	−263.65

[a] Asymptotic normal test statistics in parentheses; these statistics were obtained from the estimating information matrix.

does a simple static model with a "components-of-variance" scheme (column 8 of Table 7.3, column 5 of Table 7.4) perform any better. Dynamic models that neglect heterogeneity (column 3 of Table 7.4) overestimate labor-market turnover, whereas the static model with a conventional variance components formulation (column 5 of Table 7.4) overstates the extent of heterogeneity and the degree of intertemporal correlation. It over-predicts the number who never work during these three years and underpredicts the number who always work.

This example suggests that considerable care should be exercised in utilizing panel data to discriminate among state dependence, heterogeneity, and serial correlations. Improper control for heterogeneity can lead to erroneous parameter estimates and dramatically overstate the impact of past experience on current choices.

7.5.5.2 Household Brand Choices

Chintagunta, Kyriazidou, and Perktold (2001) use the A.C. Nielson data on yogurt purchases in Sioux Falls, South Dakota between September 17, 1986 and August 1, 1988 to study yogurt brand loyalty. They focus on the 6 oz.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Table 7.3. *(cont.)*

(4)	(5)	(6)	(7)	(8)
−2.367 (6.4)	−2.011 (3.4)	−2.37 (5.5)	−3.53 (4.6)	−1.5 (0)
−0.742 (2.6)	−0.793 (2.1)	−0.70 (2.0)	−1.42 (2.3)	−0.69 (1.2)
−0.030 (1.5)	−0.027 (1.2)	−0.03 (1.6)	−0.059 (1.3)	0.046 (11)
0.090 (0.93)	0.139 (1.5)	0.13 (1.4)	0.27 (1.1)	0.105 (0.68)
−0.124 (4.9)	−0.116 (2.2)	−0.161 (4.9)	−0.203 (3.9)	−0.160 (6.1)
0.152 (7.3)	0.095 (2.5)	0.077 (3)	0.196 (4.8)	0.105 (3.3)
-0.312×10^{-4} (5.2)	-0.207×10^{-4} (2.3)	-0.2×10^{-4} (2.6)	-0.65×10^{-4} (5.1)	-0.385×10^{-4} (20)
−0.003 (0.38)	−0.021 (0.26)	0.02 (3)	1.03 (0.14)	−0.71 (0)
—[b]	—	—	—	—
0.062 (0.38)	0.062 (3.5)	0.091 (7.0)	0.101 (5.4)	0.095 (11.0)
0.917	—	—	—	—
0.873	—	—	—	—
0.946	—	—	—	—
—	−0.942 (50)	—	—	—
—	—	0.92 (4.5)	—	0.941 (4.1)
−239.81	−243.11	−244.7	−367.3	−242.37

[b] Not estimated.

Source: Heckman (1981c, Table 3.2).

packages of the two dominant yogurt brands, Yoplait and Nordica, for the analysis. These brands account for 18.4 and 19.5 percent of yogurt purchases by weight. Only data for households that have at least two consecutive purchases of any one of the two brands are considered. This leaves 737 households and 5618 purchase occasions, out of which 2718 are for Yoplait and the remaining 2900 for Nordica. The panel is unbalanced.[20] The minimum number of purchase occasions per household is 2 and the maximum is 305. The mean number of purchase is 9.5 and the median is 5.

The model they estimate is given by

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, y_{i0}, \ldots, y_{i,t-1}, \alpha_i) = \frac{\exp(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i)}, \tag{7.5.40}$$

where $y_{it} = 1$ if household i chooses Yoplait in period t and $y_{it} = 0$ if household i chooses Nordica. The exogenous variables in $\mathbf{x}_{it}$ are the difference in the

[20] One can modify the estimator (7.5.33) by replacing T with T_i.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Table 7.4. *Comparisons of employment models using run data: Women aged 45–59 in 1968*

	(1)	(2)	(3)	(4)	(5)
Run pattern	Actual number	Number predicted from state-dependent model with heterogeneity (column 2 of Table 7.3)	Probit model that ignores heterogeneity (column 3 of Table 7.3)	Probit model that ignores heterogeneity and recent-sample state dependence (column 7 of Table 7.3)	Number predicted from Heckman–Willis model (column 8 of Table 7.3)
0,0,0	96	94.2	145.3	36.1	139.5
0,0,1	5	17.6	38.5	20.5	4.1
0,1,0	4	1.8	1.9	20.2	4.1
1,0,0	8	2.6	0.35	20.6	4.1
1,1,0	5	1.4	0.02	21.2	3.6
1,0,1	2	2.4	1.38	21.1	3.6
0,1,1	2	16.4	8.51	21.7	3.6
1,1,1	76	61.5	2.05	36.6	34.9
$\chi^{2\,c}$	—	48.5	4,419	221.8	66.3

[a] Data for 1971, 1972, and 1973, three years following the sample data, were used to estimate the model.

[b] 0 corresponds to not working; 1 corresponds to working; thus, 1,1,0 corresponds to a woman who worked the first two years of the sample and did not work in the final year.

[c] This is the standard chi-square statistic for goodness of fit. The higher the value of the statistic, the worse the fit.

Source: Heckman (1981c).

natural logarithm of the price (coefficient denoted by β_P) and the difference in the dummy variables for the two brands that describe whether the brand was displayed in the store and featured in an advertisement that week (coefficients denoted by β_D and β_F respectively). Among the many models they estimated, Table 7.5 presents the results of

1. The pooled logit model, with the lagged choice treated as exogenous assuming there are no individual specific effects (PLL)
2. The Chamberlain (1982) conditional logit approach with the lagged choice treated as exogenous (CLL)
3. The pooled logit approach with normally distributed random effects with mean μ and variance σ_α^2, with the initial choice treated as exogenous (PLL-HET)
4. The pooled logit approach with normally distributed random effects and the initial probability of choosing 1 given $(\mathbf{x}_i, \alpha_i)$ assuming at the

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

Table 7.5. *Estimates of brand choices using various approaches (standard errors in parentheses)*

Model	β_p	β_d	β_f	γ	μ_α	σ_α
CLL	−3.347	0.828	0.924	−0.068		
	(0.399)	(0.278)	(0.141)	(0.140)		
PLL	−3.049	0.853	1.392	3.458	−0.333	
	(0.249)	(0.174)	(0.091)	(0.084)	(0.102)	
PLLHET	−3.821	1.031	1.456	2.126	0.198	1.677
	(0.313)	(0.217)	(0.113)	(0.114)	(0.150)	(0.086)
PLLHETE	−4.053	0.803	1.401	1.598	0.046	1.770
	(0.274)	(0.178)	(0.115)	(0.115)	(0.133)	(0.102)
HK05	−3.477	0.261	0.782	1.223		
	(0.679)	(0.470)	(0.267)	(0.352)		
HK10	−3.128	0.248	0.759	1.198		
	(0.658)	(0.365)	(0.228)	(0.317)		
HK30	−2.644	0.289	0.724	1.192		
	(0.782)	(0.315)	(0.195)	(0.291)		
PLLHET-S[a]	−3.419	1.095	1.291	1.550	0.681	1.161
	(0.326)	(0.239)	(0.119)	(0.117)	(0.156)	(0.081)

[a] The PLLHET estimates after excluding those households that are completely loyal to one brand.

Source: Chintagunta, Kyriazidou, and Perktold (2001, Table 3).

steady state, which is approximated by

$$\frac{F(\bar{\mathbf{x}}_i'\beta + \alpha_i)}{1 - F(\bar{\mathbf{x}}_i'\beta + \gamma + \alpha_i) + F(\bar{\mathbf{x}}_i'\beta + \alpha_i)}, \tag{7.5.41}$$

where $F(a) = \exp(a)/(1 + \exp(a))$ and $\bar{\mathbf{x}}_i$ denotes the individual time series mean of $\mathbf{x}_{it}$ (PLLHETE)

5. The Honoré and Kyriazidou (2000a) approach, where $h_N = c \cdot N^{-1/5}$ with $c = 0.5$ (HK05), 1.0 (HK10), and 3.0 (HK30)

Table 7.5 reveals that almost all procedures yield statistically significant coefficients with the expected signs. An increase in the price of a brand reduces the probability of choosing the brand, and the presence of a store display or of a feature advertisement for a brand makes purchase of that brand more likely. Also, apart from CLL, all methods produce positive and statistically significant estimates for γ, that is, a previous purchase of a brand increases the probability of purchasing the same brand in the next period. The lagged choice is found to have a large positive effect in brand choice for pooled methods assuming no heterogeneity: PLL estimates of γ is 3.5. However, introducing heterogeneity lowers it substantially to 2.1 (PLLHET). The estimate of γ further drops to 1.598 (PLL-HETE) when the initial observations are treated as endogenous, and drops to about 1.2 using the Honoré–Kyriazidou estimator. Nevertheless, they do indicate that after controlling for the effects of α_i, a previous purchase of a

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

brand increases the probability of purchasing the same brand in the next period, although their impact is substantially reduced compared to the case of assuming no heterogeneity. There is also an indication of substantial heterogeneity in the sample. All methods that estimate random effects give high values for the standard deviation of the household effects, σ_α about 1.7, bearing in mind that σ_u is normalized to 1 only.

In general, the size of the estimated parameters varies considerably across estimation methods. There is also some sensitivity in the HK point estimates of all coefficients with respect to the bandwidth choice. To investigate this issue further and identify situations where the different methods are most reliable in producing point estimates, Chintagunta, Kyriazidou, and Perktold (2001) further conduct Monte Carlo studies. Their results indicate that the conditional likelihood procedures are the most robust in estimating the coefficients on the exogenous variables. However, the coefficient on the lagged dependent variable is significantly underestimated. The pooled procedures are quite sensitive to model misspecification, often yielding large biases for key economic parameters. The estimator proposed by Honoré and Kyriazidou (2000a) performs quite satisfactory despite a loss of precision because their method de facto only uses substantially smaller number of observations than other methods due to the use of the weighting scheme $K\left(\frac{\mathbf{x}_{it}-\mathbf{x}_{is}}{h_N}\right)$.

7.6 ALTERNATIVE APPROACHES FOR IDENTIFYING STATE DEPENDENCE

Section 7.5 focuses on getting consistent estimators for dynamic panel discrete choice models with individual-specific effects. If individual-specific effects are treated as random, the consistency of dynamic models requires the knowledge of the conditional distribution of individual-specific effects α_i given the T time series observations of the $K \times 1$ exogenous variables, $\mathbf{x}_{it}, \mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, $G(\alpha_i \mid \mathbf{x}_i)$, and the initial value distribution given $\mathbf{x}_i$, $P(y_{i0} \mid \mathbf{x}_i)$. If α_i is treated as a fixed constant, the consistency of the MLE requires $T \to \infty$. If T is finite, the conditions for obtaining consistent estimator of the coefficients of exogenous variables and lagged dependent variables impose severe restrictions on the observed data that only very small proportion of the sample may be utilized, if they satisfy the conditions at all. In this section, we consider alternative approaches to identify the dynamic dependence–bias reduced estimator for fixed-effects models; bounding parameters without the knowledge of $G(\alpha \mid \mathbf{x})$ and $P(y_0 \mid \mathbf{x})$ for random effects models; and approximate model.

7.6.1 Bias-Adjusted Estimator

Controlling the impact of unobserved heterogeneity in linear models are relatively straightforward (e.g., see Chapters 3 and 4). Controlling the impact of

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

unobserved heterogeneity that is correlated with explanatory variables in nonlinear models is much more difficult. When T is finite, the estimators of the parameters of interest (structural parameters) are inconsistent no matter how large N is. This inconsistency occurs because only a finite number of observations are available to estimate each individual effect α_i while the estimation of structural parameters depends on α_i. Increasing T does not necessarily fully solve this problem if N also grows with T (e.g., see Alvarez and Arellano 2003; Hahn and Newey 2004). In this section we consider methods that reduce the bias of the estimator to the order of $\frac{1}{T^2}$.

Let $\boldsymbol{\theta}$ denote the parameters of interest (structural parameters) and α_i denote the unobserved individual-specific effects (incidental parameters). Let $\hat{\boldsymbol{\theta}}_T$ denote the estimator of $\boldsymbol{\theta}$ based on NT panel data $(y_{it}, \mathbf{x}_{it})$ and $\hat{\alpha}_i$, the estimated α_i, say the fixed effects MLE (7.3.2) for static logit model ((7.3.2) and 7.3.3)) or dynamic logit model (7.5.16). In general, because of the error in the estimation of α_i when T is fixed, as $N \to \infty$, $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}_T$, where

$$\boldsymbol{\theta}_T = \boldsymbol{\theta} + \frac{B}{T} + \frac{D}{T^2} + O\left(\frac{1}{T^3}\right) \tag{7.6.1}$$

for some B and D. This bias should be small for large T. However, if N grows at the same rate as T when $T \to \infty$, the fixed-effects estimator $\hat{\boldsymbol{\theta}}$ is asymptotically biased. For $\frac{N}{T} \to c \neq 0$,

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \sqrt{NT} \cdot \frac{B}{T} + O\left(\sqrt{\frac{N}{T^3}}\right) \tag{7.6.2}$$

will have asymptotic normal distribution centered at $\sqrt{c}B$. (e.g., the fixed-effects estimator for the dynamic panel data model (4.2.3)).

Hahn and Newey (2004) suggest a jackknife estimator to reduce the bias,

$$\tilde{\theta} \equiv T\hat{\boldsymbol{\theta}} - \frac{T-1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\theta}}(t), \tag{7.6.3}$$

where $\hat{\boldsymbol{\theta}}(t)$ be the fixed effects estimator based on the subsample excluding the observations of the tth period. If $\boldsymbol{\theta}_T$ has the form (7.6.1), then the estimator $\tilde{\boldsymbol{\theta}}$ will converge in probability to

$$\begin{aligned} &(T\boldsymbol{\theta}_T - (T-1)\boldsymbol{\theta}_{T-1}) \\ &\quad = \boldsymbol{\theta} + \left(\frac{1}{T} - \frac{1}{T-1}\right) D + O\left(\frac{1}{T^2}\right) \\ &\quad = \boldsymbol{\theta} + O\left(\frac{1}{T^2}\right). \end{aligned} \tag{7.6.4}$$

Thus, the jackknife estimator reduces the bias to the order of $\frac{1}{T^2}$. However, in addition to the fact that the jackknife estimator (7.6.3) requires the estimation of

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

$(T+1)$ fixed effects estimators, the asymptotic covariance matrix of $\tilde{\boldsymbol{\theta}}$ is complicated to derive unless $(y_{it}, \mathbf{x}_{it})$ are contemporaneously and intertemporally independently distributed (over i and t).

An alternative approach is to obtain an estimated B, $\hat{B}$, then forming a bias corrected estimator

$$\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}} - \frac{\hat{B}}{T}, \tag{7.6.5}$$

(e.g. (4.7.28)). The advantage of (7.6.5) is that it reduces the bias while the formula for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^*$ remains the same as that of $\hat{\boldsymbol{\theta}}$. However, the derivation of $\hat{B}$ can be complicated.

For instance, consider the panel dynamic binary choice model of the form,

$$\begin{aligned} y_{it} &= 1(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i + u_{it} > 0), \\ i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \\ y_{i0} &\text{ observable,} \end{aligned} \tag{7.6.6}$$

where $1(A) = 1$ if event A occurs and 0 otherwise. We suppose that u_{it} is independently, identically distributed with mean 0. Then

$$\begin{aligned} E(y_{it} \mid y_{i,t-1}, \mathbf{x}_{it}, \alpha_i) &= \text{Prob}\,(y_{it} = 1 \mid y_{i,t-1}, \mathbf{x}_{it}, \alpha_i) \\ &= F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i) \\ &= F_{it}, \end{aligned} \tag{7.6.7}$$

where F is the integral of the probability distribution function of u_{it} from $-(\mathbf{x}_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)$ to ∞. When α_i is considered fixed, the log-likelihood function conditional on y_{i0} takes the form

$$\log L = \sum_{i=1}^{N}\sum_{t=1}^{T}[y_{it} \log F_{it} + (1 - y_{it}) \log (1 - F_{it})] \tag{7.6.8}$$

The MLE of $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \gamma)$ and α_i are obtained by solving the following first-order conditions simultaneously:

$$\frac{\partial \log L}{\partial \alpha_i} \Big|_{\hat{\alpha}_i} = 0, \quad i = 1, \ldots, N, \tag{7.6.9}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{7.6.10}$$

Substituting the solutions of (7.6.9) as function of $\boldsymbol{\theta}$ to (7.6.8) yields the concentrated log-likelihood function

$$\log L^* = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})), \tag{7.6.11}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

where

$$\ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = \sum_{t=1}^{T} \Big\{ y_{it} \log F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \hat{\alpha}_i(\boldsymbol{\theta})) + (1 - y_{it}) \log [1 - F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \hat{\alpha}_i(\boldsymbol{\theta}))] \Big\}.$$

Then the MLE of $\boldsymbol{\theta}$ is the solution of the following first-order conditions:

$$\frac{1}{NT} \sum_{i=1}^{N} \left[\frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\theta))}{\partial \boldsymbol{\theta}} + \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i(\boldsymbol{\theta})} \times \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{\hat{\boldsymbol{\theta}}} = \mathbf{0}. \quad (7.6.12)$$

The estimating equation (7.6.12) depends on $\hat{\alpha}_i$. When $T \to \infty$, $\hat{\alpha}_i \to \alpha_i$, the MLE of $\boldsymbol{\theta}$ is consistent. When T is finite, $\hat{\alpha}_i \neq \alpha_i$, then (7.6.12) evaluated at $\boldsymbol{\theta}$ does not converge to 0. Hence the MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is not consistent. The bias of the MLE is of order $\frac{1}{T}$. The analytical solution for $\hat{B}_T$ can be derived by taking a Taylor series expansion of (7.6.12) (e.g., see Hahn and Kuersteiner 2011).

Instead of obtaining $\hat{B}_T$ directly, Carro (2007) proposes to derive the bias corrected MLE directly by taking the Taylor series expansion of the score function (7.6.12) around α_i and evaluate it at the true value $\boldsymbol{\theta}$ yields

$$d_{\theta i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = d_{\theta i}(\boldsymbol{\theta}, \alpha_i) + d_{\theta\alpha_i i}(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i) + \frac{1}{2} d_{\theta\alpha_i\alpha_i i}(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)^2 + O_p(T^{-\frac{1}{2}}), \; i = 1, \ldots, N. \quad (7.6.13)$$

where $d_{\theta i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} + \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i(\boldsymbol{\theta}))} \cdot \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, $d_{\boldsymbol{\theta}\alpha_i i} = \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \hat{\alpha}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \alpha_i}$, $d_{\theta\alpha_i\alpha_i i} = \frac{\partial^3 \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \alpha_i \partial \alpha_i}$. Making use of McCullah (1987) asymptotic expansion for $(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)$ and $(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)^2$, Carro (2007) derives the bias-corrected estimator from the modified score function of $d_{\theta_i} = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$,

$$\sum_{i=1}^{N} d_{\theta_i}^{*} = \sum_{i=1}^{N} \Big\{ d_{\theta_i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) - \frac{1}{2} \frac{1}{d_{\alpha_i\alpha_i i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))} \Big(d_{\theta\alpha_i\alpha_i i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) + d_{\alpha_i\alpha_i\alpha_i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big) + \frac{\partial}{\partial \alpha_i} \Big(\frac{1}{E[d_{\alpha_i\alpha_i i}(\boldsymbol{\theta}, \alpha_i)]} E[d_{\theta\alpha_i i}(\boldsymbol{\theta}, \alpha_i)] \Big) \Big|_{\hat{\alpha}_i(\theta)} \Big\}_{\hat{\boldsymbol{\theta}}^{*}} = \mathbf{0}, \quad (7.6.14)$$

where $d_{\alpha_i\alpha_i i} = \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \alpha_i(\boldsymbol{\theta}))}{\partial \alpha_i^2}$ and $d_{\alpha_i\alpha_i\alpha_i} = \frac{\partial^3 \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i^3(\boldsymbol{\theta})}$. Carro (2007) shows that the bias of the modified MLE, $\hat{\boldsymbol{\theta}}^{*}$, is also of order $\left(\frac{1}{T^2}\right)$ and has the same asymptotic variance as the MLE. His Monte Carlo studies show that the bias of the modified MLE is small with $T = 8$.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

7.6.2 Bounding Parameters

When y_{i0} and α_i are treated as random, the joint likelihood of $f(\mathbf{y}_i, y_{i0} \mid \mathbf{x}_i)$ can be written in the form of conditional density of $f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i)$ times the marginal density $f(y_{i0} \mid \mathbf{x}_i)$,

$$\begin{aligned} & f(\mathbf{y}_i, y_{i0} \mid \mathbf{x}_i) \\ & \quad = \int f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) f(\mathbf{y}_{i0} \mid \mathbf{x}_i, \alpha_i) G(\alpha_i \mid \mathbf{x}_i) d\alpha_i, \\ & i = 1, \ldots, N, \end{aligned} \tag{7.6.15}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})$, and $G(\alpha_i \mid \mathbf{x}_i)$ denotes the conditional density of α_i given $\mathbf{x}_i$. For model (7.6.6) with u_{it} following a standard normal distribution, $N(0, 1)$,

$$\begin{aligned} f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) = \prod_{t=1}^{T} & [\Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)]^{y_{it}} \\ & \cdot [1 - \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)]^{1-y_{it}} \\ & i = 1, \ldots, N. \end{aligned} \tag{7.6.16}$$

If u_{it} follows a logistic distribution

$$\begin{aligned} f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) = \prod_{t=1}^{T} & \frac{\exp\left[(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)\right]^{y_{it}}}{1 + \exp\left(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)}, \\ & i = 1, \ldots, N. \end{aligned} \tag{7.6.17}$$

When $G(\alpha \mid \mathbf{x})$ and the initial distribution $P(y_0 \mid \mathbf{x})$ are known, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma)$ can be estimated by the MLE. However, $G(\alpha \mid \mathbf{x})$ and $f(y_0 \mid \mathbf{x})$ are usually unknown. Although in principle, one can still maximize (7.6.15), the usual regularity conditions for the consistency of the MLE (e.g., Kiefer and Wolfowitz 1956) is violated because $G(\alpha \mid \mathbf{x})$ is infinite dimensional (Cosslett 1981).

If $\mathbf{x}$ and α are discrete, Honoré and Tamer (2006) suggest a linear program approach to provide the bound of $\boldsymbol{\theta}$. Let $A_j = (d_{j1}, \ldots, d_{jT})$ be the $1 \times T$ sequence of binary variables d_{jt}. Let $\mathcal{A}$ denote the set of all 2^T possible sequence of 0's and 1's, A_j. Let $P(y_{i0} \mid \mathbf{x}_i, \alpha_i)$ denote the probability of $y_{i0} = 1$ given $\mathbf{x}_i$ and α_i and $f_0(\alpha, \mathbf{x})$ denote the distribution of y_{i0} given $\mathbf{x}$ and α. Then, conditional on $P(y_{i0} \mid \mathbf{x}_i, \alpha_i)$,

$$\begin{aligned} f(\mathbf{y}_i \mid \mathbf{x}_i, f_0(y_{i0} \mid \mathbf{x}_i, \boldsymbol{\alpha}_i), \alpha_i) = {} & P(y_{i0} \mid \mathbf{x}_i, \alpha_i) f(\mathbf{y}_i \mid y_{i0} = 1, \mathbf{x}_i, \alpha_i) \\ & + (1 - P(y_{i0} \mid \mathbf{x}_i, \alpha_i)) f(\mathbf{y}_i \mid y_{i0} = 0, \mathbf{x}_i, \alpha_i), \end{aligned} \tag{7.6.18}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

and

$$f(\mathbf{y}_i \mid \mathbf{x}_i, f_0(\cdot,\cdot), \boldsymbol{\theta}) = \int f(\mathbf{y}_i \mid \mathbf{x}_i, \alpha, f_0(\cdot,\cdot), \boldsymbol{\theta}) dG(\alpha \mid \mathbf{x}_i). \tag{7.6.19}$$

Let $\pi(A \mid \mathbf{x}, \alpha; f_0(\cdot,\cdot), \boldsymbol{\theta})$ and $P(A \mid \mathbf{x})$ be the probability of an event A in $\mathcal{A}$ given $(\mathbf{x}, \alpha)$ predicted by the model and the probability of an event A occurs given $\mathbf{x}$, respectively. Then $\pi(A \mid \mathbf{x}, f_0(\cdot,\cdot), \boldsymbol{\theta}) = \int \pi(A \mid \mathbf{x}, \alpha; f_0(\cdot,\cdot), \boldsymbol{\theta}) dG(\alpha \mid \mathbf{x})$. Define the set of $(f_0(\cdot,\cdot), \boldsymbol{\theta})$ that is consistent with a particular data-generating process with probabilities $\mathcal{P}(\mathcal{A} \mid \mathbf{x})$ as

$$\Psi = \Big\{ (f_0(\cdot,\cdot), \boldsymbol{\theta}) : P[\pi(\mathcal{A} \mid \mathbf{x}; f_0(\cdot,\cdot), \boldsymbol{\theta}) = P(\mathcal{A} \mid \mathbf{x})] = 1 \Big\}. \tag{7.6.20}$$

Then the bound of $\boldsymbol{\theta}$ is given by

$$\ominus = \Big\{ \boldsymbol{\theta} : \exists f_0(\cdot,\cdot) \text{ such that } P[\pi(\mathcal{A} \mid \mathbf{x}; f_0(\cdot,\cdot), \boldsymbol{\theta}) = P(\mathcal{A} \mid \mathbf{x})] = 1) \Big\}. \tag{7.6.21}$$

Suppose α has a discrete distribution with a known maximum number of points of support, M. The points of support are denoted by a_m and the probability of $\alpha_i = a_m$ given $\mathbf{x}$ denoted by ρ_{mx}. Then

$$\begin{aligned} &\pi(\mathcal{A} \mid f_0(\cdot,\cdot), \mathbf{x}, \boldsymbol{\theta}) \\ &= \sum_{m=1}^{M} \rho_{mx} \Big[f_0(a_m, \mathbf{x}) \pi(\mathcal{A} \mid y_0 = 1, \boldsymbol{\theta}, \mathbf{x}; a_m) \\ &\quad + (1 - f_0(a_m, \mathbf{x})) \pi(\mathcal{A} \mid y_0 = 0, \boldsymbol{\theta}, \mathbf{x}; a_m) \Big] \\ &= \sum_{m=1}^{M} z_{mx} \pi(\mathcal{A} \mid y_0 = 1, \mathbf{x}, \boldsymbol{\theta}; a_m) + \sum_{m=1}^{M} z_{M+m,x} \pi(\mathcal{A} \mid y_0 = 0, \mathbf{x}, \boldsymbol{\theta}; a_m), \end{aligned} \tag{7.6.22}$$

where $z_{mx} = \rho_{mx} f_0(a_m, x)$ and $z_{M+m,x} = \rho_{mx}[1 - f_0(a_m, \mathbf{x})]$ for $m = 1, \ldots, M$. The identified set $\ominus$, consists of the value of $\boldsymbol{\theta}$ for which the following equations have a solution for $\{z_{mx}\}_{m=1}^{2M}$:

$$\sum_{m=1}^{M} z_{mx} \pi(A \mid y_0 = 1, \mathbf{x}, \boldsymbol{\theta}; a_m) + \sum_{m=1}^{M} z_{M+m,x} \pi(A \mid y_0 = 0, \mathbf{x}, \boldsymbol{\theta}; a_m) = P(A \mid \mathbf{x}), \tag{7.6.23}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

and for all $A \in \mathcal{A}$,

$$\sum_{m=1}^{2M} z_{mx} = 1, z_{mx} \geq 0. \tag{7.6.24}$$

Equation's (7.6.23) and (7.6.24) have exactly the same structure as the constraints in a linear programming problem, so checking whether a particular $\boldsymbol{\theta}$ belongs to $\ominus$ can be done in the same way that checks for a feasible solution in a linear programming problem provided $P(A \mid \mathbf{x})$ can be consistently estimated. Therefore, Honoré and Tamer (2006) suggest bounding $\boldsymbol{\theta}$ by considering the linear programming problem:

$$\begin{aligned} &\text{maximize} \quad \sum_j -v_{jx} \\ &\{z_{mx}, \{v_{jx}\}\} \end{aligned} \tag{7.6.25}$$

where

$$v_{jx} = P(A_j \mid \mathbf{x}) - \sum_{m=1}^{M} z_{mx} \pi(A_j \mid y_{i0} = 1, \mathbf{x}, \boldsymbol{\theta}; a_m)$$

$$- \sum_{m=1}^{M} z_{M+m,x} \pi(A_j \mid y_{i0} = 0, \mathbf{x}, \boldsymbol{\theta}; a_m)$$

$$\text{for all } A_j \in \mathcal{A}, \ j = 1, \ldots, 2^T, \tag{7.6.26}$$

$$1 - \sum_{m=1}^{2M} z_{mx} = v_{0x}, \tag{7.6.27}$$

$$z_{mx} \geq 0, \tag{7.6.28}$$

$$v_{jx} \geq 0. \tag{7.6.29}$$

The optimal function value for (7.6.25) is 0 if and only if all $v_{jx} = 0$, that is, if a solution exists to (7.6.23) and (7.6.24). If (7.6.23) and (7.6.24) do not have a solution, the maximum function value in (7.6.25) is negative. Following Manski and Tamer (2002) it can be shown that a consistent estimator of the identified region can be constructed by checking whether, for a given $\boldsymbol{\theta}$, the sample objective function is within ϵ of the maximum value of 0 where $P(A)$ is substituted by its consistent estimator. Because $\mathbf{x}$ is discrete, one can mimic this argument for each value in the support of $\mathbf{x}_i$ which will then contribute a set of constraints to the linear programming problem.

7.6.3 Approximate Model

The dynamic logit model (7.5.24) implies that the conditional distribution of a sequence of response variables, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ given α_i, $\mathbf{x}_i$, and y_{i0}, can

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

be expressed as

$$P(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}) = \frac{\exp\left(\alpha_i \sum_{t=1}^{T} y_{it} + \sum_{t=1}^{T} y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}) + y_{i*}\gamma\right)}{\sum_{t=1}^{T}[1 + \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)]}, i = 1, \ldots, N, \tag{7.6.30}$$

where $y_{i*} = \sum_{t=1}^{T} y_{i,t-1} y_{it}$.

The Honoré and Kyriazidou (2000a) conditional approach discussed in Section 7.5 requires a specification of a suitable kernel function and the bandwidth parameters to weigh the response configuration of each subject in the sample on the basis of exogenous explanatory variables in which only exogenous variables are close to each other receiving large weights, implying a substantial reduction of the rate of convergence of the estimator to the true parameter value. Moreover, conditional on certain configurations leads to a response function in terms of time changes of the covariates, implying the exclusion of time-invariant variables. Noting that the dynamic logit model (7.6.30) implies that the conditional log-odds ratio between $(y_{it}, y_{i,t-1})$ equals to

$$\log \frac{P(y_{it} = 0 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 0) \cdot P(y_{it} = 1 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 1)}{P(y_{it} = 0 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 1) \cdot P(y_{it} = 1 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 0)} = \gamma, \tag{7.6.31}$$

Bartolucci and Nigro (2010) suggest using the Cox (1972) quadratic exponential model to approximate (7.6.30),[21]

$$P^*(\mathbf{y}_i \mid \mathbf{x}_i, y_{i0}, \delta_i) = \frac{\exp\left[\delta_i(\sum_{t=1}^{T} y_{it}) + \sum_{t=1}^{T} y_{it}(\mathbf{x}_{it}'\boldsymbol{\phi}_1) + y_{iT}(\psi + \mathbf{x}_{iT}'\boldsymbol{\phi}_2) + y_{i*}\tau\right]}{\sum_{d_i} \exp[\delta_i(\sum_{t=1}^{T} d_{ijt}) + \sum_{t=1}^{T} d_{ijt}(\mathbf{x}_{it}'\boldsymbol{\phi}_1) + d_{ijT}(\psi + \mathbf{x}_{iT}'\boldsymbol{\phi}_2) + d_{ij}^*\tau]} \tag{7.6.32}$$

where $\mathbf{d}_{ij} = (d_{ij1}, \ldots, d_{ijT})$ denote the jth possible binary response sequence, $\sum_{d_i}$ denotes the sum over all possible response sequence of $\mathbf{d}_{ij}$, such that $\sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it}$, and $d_{ij}^* = d_{ij1} y_{io} + \sum_{t=1}^{T} d_{ijt} d_{ij,t-1}$. Model (7.6.32) implies that

$$P^*(y_{it} \mid \mathbf{x}_i, \delta_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp\{y_{it}[\delta_i + \mathbf{x}_{it}'\boldsymbol{\phi}_1 + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i)]\}}{1 + \exp\left[\delta_i + \mathbf{x}_{it}'\boldsymbol{\phi}_1 + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i)\right]}, \tag{7.6.33}$$

[21] To differentiate the approximate model from the dynamic logit model (7.6.30), we use δ_i to represent the individual-specific effects and $\boldsymbol{\phi}_1$ to represent the coefficients of $\mathbf{x}_{it}$ in the approximate model (7.6.32).

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

where, for $t < T$,

$$\begin{aligned} e_t^*(\delta_i, \mathbf{x}_i) &= \log \frac{1 + \exp\left[\delta_i + \mathbf{x}'_{i,t+1}\boldsymbol{\phi}_1 + e_{t+1}^*(\delta_i, \mathbf{x}_i) + \tau\right]}{1 + \exp\left[\delta_i + \mathbf{x}'_{i,t+1}\boldsymbol{\phi}_1 + e_{t+1}^*(\delta_i, \mathbf{x}_i)\right]} \\ &= \log \frac{P(y_{i,t+1} = 0 \mid \delta_i, \mathbf{x}_i, y_{it} = 0)}{P(y_{i,t+1} = 0 \mid \delta_i, \mathbf{x}_i, y_{it} = 1)}. \end{aligned} \tag{7.6.34}$$

The corrections term (7.6.34) depends on future covariates. For the last period, it is approximated by

$$e_T^*(\delta_i, \mathbf{x}_i) = \psi + \mathbf{x}'_{iT}\boldsymbol{\phi}_2. \tag{7.6.35}$$

Model (7.6.33) may be viewed as a latent response model of the form

$$y_{it}^* = \mathbf{x}'_{it}\boldsymbol{\phi}_1 + \delta_i + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i) + \eta_{it}, \tag{7.6.36}$$

with logistically distributed stochastic term η_{it}. The correction term $e_t^*(\delta_i, \mathbf{x}_i)$ may be interpreted as a measure of the effect of the present choice y_{it} on the expected utility (or propensity) at period $(t + 1)$. The parameter τ for the state dependence is the log-odds ratio between any pairs of variables $(y_{i,t-1}, y_{it})$, conditional on all the other response variables or marginal with respect to these variables.

The difference between the approximate model (7.6.32) and the dynamic logit model (7.6.30) is in the denominator. The former does not depend on the actual sequence $\mathbf{y}_i$, while the latter does. The advantage of model (7.6.32) or (7.6.33) is that the parameters for the unobserved heterogeneity, δ_i, can be eliminated by conditioning on the sum of response variables over time just like the static logit model (7.3.14). When y_{i0} are observable, the structural parameters can be estimated by the conditional maximum likelihood estimator as discussed in (7.3.21) when $T \geq 2$.

The relations between (7.6.32) and (7.6.30) can be seen through a Taylor series expansion of the nonlinear term of the logarithm of the dynamic logit model (7.6.30) at $\alpha_i = \tilde{\alpha}_i$, $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\gamma = 0$,

$$\begin{aligned} &\sum_{t=1}^{T} \log\left[1 + \exp\left(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)\right] \\ &\quad \simeq \sum_{t=1}^{T} \{ \log\left[1 + \exp\left(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}} + \tilde{\alpha}_i\right)\right] + \tilde{q}_{it} \\ &\qquad \cdot \left[\mathbf{x}'_{it}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + (\alpha_i - \tilde{\alpha}_i)\right]\} + \tilde{q}_{i1} y_{i0}\gamma + \sum_{t=1}^{T} \tilde{q}_{it} y_{i,t-1}\gamma, \end{aligned} \tag{7.6.37}$$

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

where

$$\tilde{q}_{it} = \frac{\exp(\tilde{\alpha}_i + \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}})}{1 + \exp(\tilde{\alpha}_i + \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}})}. \tag{7.6.38}$$

Substituting (7.6.37) into the logarithm of (7.6.30) and renormalizing the exponential of the resulting expression yields the approximate model for (7.6.30) as

$$\begin{aligned} &P^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}) \\ &= \frac{\exp(\alpha_i(\sum_{t=1}^T y_{it}) + \sum_{t=1}^T y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}) - \sum_{t=2}^T \tilde{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)}{\sum_{d_i} \exp[\alpha_i(\sum_{t=1}^T d_{ijt}) + \sum_{t=1}^T d_{ijt}(\mathbf{x}_{it}'\boldsymbol{\beta}) - (\sum_{t=2}^T \tilde{q}_{it} d_{i,t-1})\gamma + d_{ij*}\gamma)]}, \\ &i = 1, \ldots, N. \end{aligned} \tag{7.6.39}$$

When γ is indeed equal to 0, the true model and the approximating model coincide. Both become the static logit model (7.3.13). The approximating model (7.6.32) or (7.6.39) implies that the conditional logit of y_{it} given $\mathbf{x}_i'$, α_i and $y_{i0}, \ldots, y_{i,t-1}$,is equal to

$$\begin{aligned} &\log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1})}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1})} \\ &= \begin{cases} \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i + e_t(\alpha_i, \mathbf{x}_i) - \tilde{q}_{i,t+1}\gamma, & \text{if } t < T, \\ \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i, & \text{if } t = T, \end{cases} \end{aligned} \tag{7.6.40}$$

where

$$\begin{aligned} e_t(\alpha_i, \mathbf{x}_i) &= \log \frac{P^*(y_{i,t+1} = 0 \mid \mathbf{x}_i, \alpha_i, y_{it} = 0)}{P^*(y_{i,t+1} = 0 \mid \mathbf{x}_i, \alpha_i, y_{it} = 1)} \\ &= \bar{q}_{i,t+1}\gamma. \end{aligned} \tag{7.6.41}$$

Equation (7.6.40) implies that

$$\log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 1)}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 1)} - \log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i\, a_i, y_{i,t-1} = 0)}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 0)} = \gamma. \tag{7.6.42}$$

Just like the dynamic logit model, the approximating model implies y_{it} is conditionally independent of $y_{i0}, \ldots, y_{i,t-2}$ given $\mathbf{x}_i, \alpha_i$ and $y_{i,t-1}$ for $t = 2, \ldots, T$ and is conditionally independent of $y_{i0}, \ldots, y_{i,t-2}, y_{i,t+2}, \ldots, y_{iT}$ given $\mathbf{x}_i, \alpha_i, y_{i,t-1}$ and $y_{i,t+1}$ for $t = 2, \ldots, T-1$. However, it has the advantage that the minimum sufficient statistics for α_i is now $\sum_{t=1}^T y_{it}$. Hence the

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press

conditional distribution of $\mathbf{y}_i$ given $\sum_{t=1}^{T} y_{it}$, where $0 < \sum_{t=1}^{T} y_{it} < T$,

$$P^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}, \sum_{t=1}^{T} y_{it})$$

$$= \frac{\exp\left\{\sum_{t=2}^{T} y_{it}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} - \sum_{t=2}^{T} \tilde{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)\right\}}{\sum_{d_i} \exp\{\sum_{t=2}^{T} d_{ijt}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} - \sum_{t=2}^{T} \tilde{q}_{it} d_{ij,t-1}\gamma + d_{ij*}\gamma)},$$

$$i = 1, \ldots, N. \tag{7.6.43}$$

To obtain the pseudo-conditional MLE of the pseudo-likelihood function (7.6.43). Bartolucci and Nigro (2010) suggest first assuming there was no state dependence ($\gamma = 0$) and maximizing the conditional log-likelihood of the static logit model (7.3.21) for those i where $0 < \sum_{t=1}^{T} y_{it} < T$ to obtain a preliminary estimate $\tilde{\boldsymbol{\beta}}$. Then substituting $\tilde{\boldsymbol{\beta}}$ into (7.6.42) to obtain the revised pseudo-conditional MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ through the Newton–Raphson iterative procedure. Their Monte Carlo studies show that the pseudo-conditional MLE has a very low bias for data generated by a dynamic logit model.

https://doi.org/10.1017/CBO9781139839327.010 Published online by Cambridge University Press