4 The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process

James J. Heckman

4.1 Introduction

This chapter considers two problems: the first, and most important, is the problem of inital conditions that arises in estimating a discrete time-discrete data stochastic process; the second problem considered is the problem of incidental parameters that besets one potentially attractive solution to the problem of initial conditions.

Before parameters generating a stochastic process with dependence among time-ordered outcomes can be estimated, the process must somehow be initialized. In much applied work in the social sciences, this problem is treated casually. Two assumptions are typically invoked: either the initial conditions or relevant presample history of the process are assumed to be truly exogenous variables, or else the process is assumed to be in equilibrium.

The first assumption is valid only if the disturbances that generate the process are serially independent or if a genuinely new process is fortuitously observed at the beginning of the sample at the analyst's disposal. If the process has been in operation prior to the time it is sampled, and if the disturbances of the model are serially dependent, the initial conditions are not exogenous variables. Treating them as exogenous variables results in inconsistent parameter estimates. The confluence of serial dependence in unobservables and state dependence in the process results in an important and neglected problem that is considered in this chapter.

The second assumption—initial stationarity of the process—does lead to a tractable solution to the problem. But this assumption is unattractive in many applications, especially when time-varying exogenous variables drive the stochastic process.

This chapter proposes exact and approximate solutions to the problem of initial conditions that arises in the class of discrete-time, discrete-data stochastic processes considered in chapter 3. Limited Monte Carlo

This research was supported by NSF Grant SOC77-27136 and Grant 10-P-90748/9-01 from the Social Security Administration. I have benefited from discussions with Tom MaCurdy and Guilherme Sedlacek on the material in this chapter. I am indebted to Gary Chamberlain for the reference to the dissertation of E. B. Andersen which guided some of my thinking and to my colleague R. D. Bock for other valuable references cited. Ralph Shnelvar and Guilherme Sedlacek gave invaluable assistance in programming the Monte Carlo studies reported here. I am solely responsible for any errors.

evidence is presented on the performance of certain estimators that are simple to use.

One potentially attractive solution to the problem of initial conditions that does not require a stationarity assumption is based on the fixed effect probit estimator proposed in chapter 3, section 3.6. However, this solution gives rise to another problem also discussed in this chapter: the problem of incidental parameters first considered by Neyman and Scott (1948).

For any panel of finite length estimators of individual fixed effects are necessarily inconsistent. As Neyman and Scott have demonstrated, the inconsistency in estimating fixed effects does not necessarily give rise to inconsistency for estimators of the structural parameters of interest, provided that estimators for these parameters can be derived that do not depend on the incidental parameters (e.g., first difference estimators in linear regression). However, it is not always possible to derive such estimators, and in such a case the inconsistency in the estimator for the fixed effect is transmitted to the estimator for the structural parameters. In particular the fixed effect probit estimator suffers from this defect.

All empirical samples are of finite size. All estimators based on necessarily finite samples are inconsistent. The issue here, as always, is whether or not the asymptotic theory provides a good guide in practical work.

In this chapter a limited set of Monte Carlo sampling experiments is conducted as a first step toward evaluating the properties of the fixed effect estimator. Two main conclusions emerge from these experiments. First, for a panel of length eight the inconsistency for the fixed effect estimator is found to be small for a multivariate probit model with strictly exogenous variables. This conclusion is in general agreement with results from a set of Monte Carlo experiments for a fixed effect logit model performed by Wright and Douglas (1976). Second, for a panel of length eight the inconsistency for the fixed effect estimator is found to be disturbingly large for a multivariate probit model with fixed effect that generates a discrete data first-order Markov process.

Additional limited Monte Carlo evidence is presented on the performance of certain alternative simple approximate procedures for the solution of the problem of initial conditions that do not rely on the fixed effect probit model. These procedures are found to produce more attractive estimates than those derived from the multivariate probit model with fixed effect.

This chapter is in four sections. In sections 4.2 and 4.3 the problem of initial conditions and some solutions to it are discussed in the specific context of a first-order Markov model generated from the latent variable models presented in chapter 3. In sections 4.4 and 4.5 results from some limited Monte Carlo sampling experiments are presented. The chapter concludes with a brief summary.

4.2 The Problem of Initial Conditions and Some Formal Solutions

In order to focus the discussion on the essential aspects of the problem of initial conditions and its solution, consider the estimation of a discrete data first-order Markov process generated as a special case of the general model presented in chapter 3. For present purposes few new points arise in a discussion of the general model, and an analysis of the Markov model, which is widely used in applied work, is of interest in its own right. For simplicity exogenous variables are initially assumed to be absent, and it is further assumed that the disturbance that generates the process has a components of variance structure.

The process is defined by dichotomization of latent variable $Y(i, t)$:

$$\begin{aligned}
&Y(i,t) = \beta_0 + \gamma d(i,t-1) + \varepsilon(i,t), \quad i = 1, \ldots, I, t = 1, \ldots, T,\\
&Y(i,t) \geq 0 \quad \text{iff } d(i,t) = 1,\\
&Y(i,t) < 0 \quad \text{iff } d(i,t) = 0,\\
&\varepsilon(i,t) = \tau(i) + U(i,t),
\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}
E(U(i,t)) = 0 = E(\tau(i)), E(\tau(i)^2) = \sigma_\tau^2,\\
E(U(i,t)^2) = \sigma_U^2 = 1, E(U(i,t)U(i',t'')) = 0, t \neq t'',\\
E(\tau(i)U(i',t'')) = 0, \text{for all } i, t', \text{and } t'',\\
d(i, 0) = \text{a fixed nonstochastic constant for individual } i.
\end{aligned}$$

Assuming that $U(i, t)$ is normally distributed, an inessential assumption in the present context, the transition probability for individual i at time t given $\tau(i)$ is

$$\begin{aligned}
&\text{Prob}[d(i,t) \mid d(i,t-1), \tau(i)]\\
&\qquad = \Phi\{[\beta_0 + \gamma d(i,t-1) + \tau(i)] \cdot [2d(i,t) - 1]\},
\end{aligned} \tag{4.2}$$

where Φ is the unit normal cumulative distribution function.

The marginal probability of $d(i, J)$, $T \geq J$ given $\tau(i)$ is

$$P[d(i,J) \mid \tau(i)] = \left(\sum_{d(i,J-1)=0}^{1} \Phi\{[\beta_0 + \gamma d(i,J-1) + \tau(i)][2d(i,J) - 1]\} \right) \cdot \left(\sum_{d(i,J-2)=0}^{1} \Phi([\beta_0 + \gamma d(i,J-2) + \tau(i)][2d(i,J-1) - 1]\} \right) \cdots \left(\sum_{d(i,1)=0}^{1} \Phi\{[\beta_0 + \gamma d(i,0) + \tau(i)][2d(i,1) - 1]\} \right). \tag{4.3}$$

In other words, it is the sum of the probabilities of all possible sequences of events prior to J that leads to a specific value for $d(i, J)$.

The process is ergodic for bounded γ, β_0, and $\tau(i)$. The limiting marginal probability for the state $d(i, t) = 1$ for all t (assuming an infinite past) is

$$\Pi_1(\tau(i)) = \frac{\Phi(\beta_0 + \tau(i))}{1 - \Phi(\beta_0 + \gamma + \tau(i)) + \Phi(\beta_0 + \tau(i))}, \tag{4.4}$$

and the limiting probability for state 0 is $\Pi_0(\tau(i)) = 1 - \Pi_1(\tau(i))$.[1] Π_1 and Π_0 are the equilibrium population proportions in state 1 and state 0, respectively. Note that, if the transition probabilities are of the probit functional form, the limiting probabilities are not of that form.

Further, if the process is in equilibrium,

$$P[d(i,J) \mid \tau(i)] = \Pi_1(\tau(i))^{d(i,J)} \Pi_0(\tau(i))^{1-d(i,J)},$$

so that it is possible to write a closed form expression for the probability.

The likelihood function for a sample of T observations per person, given nonstochastic initial condition $d(i, 0)$, is

$$\mathscr{L} = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\{[\beta_0 + \gamma d(i,t-1) + \tau][2d(i,t) - 1]\} f(\tau) d\tau, \tag{4.5}$$

where $f(\tau)$ is the population density of τ. It is not necessary to assume that τ is normally distributed, although it is convenient to do so, and for that reason this assumption is maintained for the rest of the chapter.

[1] The equilibrium probabilities are obtained by solving the equation for equilibrium $\Pi_1(\tau(i)) = \Phi(\beta_0 + \gamma + \tau(i))\Pi_1(\tau(i)) + \Phi(\beta_0 + \tau(i))\Pi_0(\tau(i))$ and using the fact that $\Pi_0(\tau(i)) = 1 - \Pi_1(\tau(i))$; see, e.g., Karlin and Taylor (1975).

Maximum likelihood estimators of β_0, γ, and the variance of τ are consistent if $T \to \infty$ and $I \to \infty$ or just $I \to \infty$. If the $\tau(i)$ are treated as parameters rather than integrated out, maximum likelihood estimators of β_0, γ, and $\tau(i)$ are consistent as $T \to \infty$, whether or not I is fixed or tends to infinity.

Thus far it has been assumed that the analyst has access to the entire history of the process. Suppose, however, that the analyst only has access to the last $T - J$ observations on the process, so that he knows $d(i, t')$, $t' = J, \ldots, T$. In this case the initial state for individual i, $d(i, J)$, is not fixed or exogenous. It is determined by the process generating the panel sample. Unless $f(\tau)$ is degenerate, random variable $d(i, J)$ is not exogenous and is in fact stochastically dependent on τ.[2] This gives rise to the problem of initial conditions.

The sample conditional likelihood function for $d(i, T), \ldots, d(i, J + 1)$ given $d(i, J)$ is

$$\mathscr{L} = \prod_{i=1}^{I} \left[\left[\int_{-\infty}^{\infty} \prod_{t=J+1}^{T} \Phi\{[\beta_0 + \gamma d(i,t-1) + \tau][2d(i,t) - 1]\} \right. \right.$$

$$\left. \left. \cdot P[d(i,J) \mid \tau] f(\tau) d\tau \right] \Big/ \int_{-\infty}^{\infty} P[d(i,J) \mid \tau] f(\tau) d\tau \right], \tag{4.6}$$

where the term inside the large brackets is the conditional probability of $d(i, T), \ldots, d(i, J + 1)$ given $d(i, J)$ for observation i. The sample likelihood for $d(i, T), \ldots, d(i, J)$ is

$$\mathscr{L} = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{t=J+1}^{T} \Phi\{[\beta_0 + \gamma d(i,t-1) + \tau][2d(i,t) - 1]\}$$

$$\cdot P[d(i,J) \mid \tau] f(\tau) d\tau. \tag{4.7}$$

[2] In the case of a degenerate $f(\tau)$ distribution, maximum likelihood estimators of the parameters of the model are consistent. Amemiya (1978) proves this for a logit first-order Markov case, and his proofs carry over fully to the probit case considered in this chapter. For a discussion about estimating a Markov model without serially dependent unobservable variables, see Anderson and Goodman (1957) and Anderson (1976).

Maximizing either the conditional likelihood function (4.6) or the unconditional likelihood function (4.7) with respect to parameters β_0, γ, and σ_τ^2 generates consistent parameter estimators that are asymptotically normally distributed if $I \rightarrow \infty$ and $T \rightarrow \infty$ or just $I \rightarrow \infty$ (provided $T \geq 2$). Recall from equation (4.3) that $P[d(i, J) \mid \tau]$ depends on β_0, γ, and $d(i, 0)$.

Unless $f(\tau)$ is degenerate, so that there is no serial dependence in the disturbance of the model, maximizing the likelihood function treats $d(i, J)$ as a fixed nonstochastic constant, that is, the function

$$\mathscr{L} = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{t=J+1}^{T} \Phi\{[\beta_0 + \gamma d(i, t-1) + \tau][2d(i,t) - 1]\} f(\tau) d\tau \quad (4.8)$$

leads to inconsistent estimators for β_0, γ, and σ_τ^2.[3] Intuitively this is so because $d(i, J)$ is stochastically dependent on τ.

Maximizing likelihood (4.6) or (4.7) is a computationally forbidding task, even in the case where initial stationarity of the process can be assumed, so that

$$P[d(i,J) \mid \tau(i)] = \Pi_2 (\tau(i))^{d(i,J)} \Pi_0 (\tau(i))^{1-d(i,J)}.$$

Further, without the assumption of stationarity it is necessary to determine $d(i, 0)$, and this information is hard to come by, although in certain contexts it is plausible to assume a specific value for this variable. These difficulties become more pronounced when exogenous variables are added to the model, so that β_0 is replaced by $\mathbf{Z}(i, t)\boldsymbol{\beta}$, where the $\mathbf{Z}(i, t)$, $i = 1, \ldots, I$, $t = 1, \ldots, T$, are bounded exogenous variables. In this case two additional problems arise.

First, it is generally untenable to assume stationarity for the marginal probabilities $P[d(i, t) \mid \tau(i)]$, $t = 1, \ldots, T$, without stringent additional restrictions on the process generating the exogenous variables, such as an assumption that the exogenous variables are generated by a stationary stochastic process—an assumption that excludes time and age trends. Second, the analyst typically will not know the values of the relevant exogenous variables in the presample period $t = 1, \ldots, J - 1$. (If these values are known, they may be substituted into the expression for$P[d(i, t) \mid \tau(i)]$ given in (4.3), replacing β_0 by $\mathbf{Z}(i, t)\boldsymbol{\beta}$.)

[3] The proof of this assertion is straightforward and amounts to showing that the expectations of the partials of the log likelihood (equation 4.8) with respect to the structural parameters do not vanish at the true parameter value.

If the presample exogenous variables are not available to the analyst, in principle the parameters of the distribution of the missing data can be estimated, provided that such a distribution exists. This procedure is a straightforward application of the work of Kiefer and Wolfowitz (1956). Thus, if the joint distribution of the presample values of $\mathbf{Z}(i, t)$, $t = 1, \ldots, J - 1$, is a finite parameter distribution $g[\mathbf{Z}(1), \ldots, \mathbf{Z}(J-1)]$, and $P^*[d(i, J) \mid \tau(i), \mathbf{Z}(1), \ldots, \mathbf{Z}(J)]$ is defined as the marginal probability of $d(i, J)$ conditional on $\mathbf{Z}(i, t)$, $t = 1, \ldots, J$, and $\tau(i)$ (obtained by substituting $\mathbf{Z}(i, t)\boldsymbol{\beta}$ for β_0 in equation 4.3), one can define

$$\begin{aligned} &P[d(i, J) \mid \tau(i), \mathbf{Z}(J)] \\ &\quad = \int \ldots \int P^*[d(i, J) \mid \tau(i), \mathbf{Z}(1), \ldots, \mathbf{Z}(J)] \\ &\quad \cdot g[\mathbf{Z}(1), \ldots, \mathbf{Z}(J-1)] d\mathbf{Z}(1), \ldots, d\mathbf{Z}(J-1), \end{aligned}$$

and with this modification likelihood functions (4.6) and (4.7) remain valid. Those likelihood functions, suitably modified, can be maximized with respect to the structural parameters of the g distribution.[4] Note that in general due to the fact that the regressors enter $P^*[d(i, J) \mid \tau(i), \mathbf{Z}(1), \ldots, \mathbf{Z}(J)]$ in a nonlinear fashion, it is not correct to use estimated sample means to replace the missing values of the exogenous variables.

This procedure is not as complicated as it appears to be. If the analyst has access to other data from which it is possible to estimate g consistently, he can use the estimated distribution to form $P[d(i, J) \mid \tau(i), \mathbf{Z}(1), \ldots, \mathbf{Z}(J)]$ for each set of values of β, γ, and $\tau(i)$. Therefore it is not necessary to estimate the parameters of the g distribution simultaneously with the structural parameters of interest.

4.3 Simpler Solutions and the Problem of Incidental Parameters

The preceding section presents the problem of initial conditions and sketches some formal solutions to it. The solutions presented there are somewhat computationally forbidding. In this section some alternative, simply computed estimators are considered.

Following a suggestion made by Mundlak (1978) for a linear regression model, it is possible to estimate a model with a components of variance structure conditional on error component $\tau(i)$ and to estimate the $\tau(i)$, $i = 1, \ldots, I$. The advantage of this approach is especially clear in the

[4] Kiefer and Wolfowitz (1956) also consider a case of nonparametric estimation of the g function.

context of estimating conditional likelihood function (4.6). Treating $\tau(i)$ as a parameter, the conditional likelihood function for $d(i, T), \ldots, d(i, J+1)$ given $d(i, J)$ is

$$\mathscr{L} = \prod_{i=1}^{I} \prod_{t=J+1}^{T} \frac{\Phi\{[\beta_0 + \gamma d(i,t-1) + \tau(i)][2d(i,t) - 1]\} P(d(i,J) \mid \tau(i))}{P(d(i,J) \mid \tau(i))}$$

$$\mathscr{L} = \prod_{i=1}^{I} \prod_{t=J+1}^{T} \Phi\{[\beta_0 + \gamma d(i,t-1) + \tau(i)][2d(i,t) - 1]\}. \qquad (4.9)$$

This model is the discrete data analogue of the linear regression model with fixed effects, and it is very easy to compute.[5] Maximizing function (4.9) with respect to $\tau(i), i = 1, \ldots, I, \beta_0$, and γ, as $T \rightarrow \infty$, these parameters are consistently estimated.[6] For further discussion of this model see chapter 3, section 3.6.

The principal advantage of this procedure is that presample information about the process is not required to estimate the structural parameters. In a model with exogenous variables, the analyst can avoid estimating the distribution of missing data and computing the marginal probability $P[d(i, J) \mid \tau(i)]$. Another advantage of this procedure is that it is not necessary to assume an arbitrary distribution of the τ to estimate the model.[7]

There are two problems with the fixed effect model.[8] First, for any observation i that does not change state over the sample period the estimated fixed effect is $\pm\infty$. A fixed effect can be chosen that perfectly explains the data for such observations. For those observations for which $\Sigma d(i, t) = T - J$, $\tau(i)$ is estimated to be ∞. For those observations for which $\Sigma d(i, t) = 0$, $\tau(i)$ is estimated to be $-\infty$. The effective sample for estimating parameters is the subsample of individuals who change state. While this may be intuitively displeasing, because it apparently manufac-

[5] A copy of a computer program that computes this model is available at cost.

[6] Note that β_0 may be absorbed into the estimated fixed effects and recovered by invoking the requirement that the mean of the fixed effects is zero.

[7] As discussed in note 3, Kiefer and Wolfowitz (1956) suggest nonparametric estimation of the density of τ. In most work in econometrics arbitrary parametric schemes are imposed. Estimating the fixed effect permits one to estimate the density of τ. This procedure is not the one proposed by Kiefer and Wolfowitz.

[8] A third disadvantage is that for large $T(> 3)$. the fixed effect estimator is not a conditional version of a model with arbitrary serial correlation in the disturbances. An assumption of a components of variance error structure is required to justify the method.

tures a form of small sample selection bias, as $T \rightarrow \infty$ this problem becomes unimportant.

Second, for fixed T, estimates of structural coefficients are inconsistent. This conclusion follows from the analyses of Neyman and Scott (1948) and Andersen (1973, pp. 68–78). Andersen explicitly considers a fixed effect logit model. The inconsistency of the maximum likelihood estimator for the structural parameters arises for the following reason. Estimators of $\tau(i)$ are necessarily inconsistent. Since the roots of the likelihood equation involve the joint solution of structural parameters and fixed effects, the inconsistency of the estimator for the fixed effects is transmitted to the estimator for the structural parameters.[9]

For these reasons the fixed effect scheme appears to be unattractive. However, the analysis of Andersen's model (as it applies to the multivariate

[9] Using Andersen's example (1973, pp. 68–71), it is possible to examine these difficulties more closely. Andersen assumes $\gamma = 0$ and lets $T = 2$. Define a dummy $\eta(t)$ that equals zero in period 1 and one in period 2. The likelihood function is thus

$$\mathscr{L} = \prod_{i=1}^{I} \prod_{t=1}^{2} \Phi\{[\tau(i) + \beta\eta(t)](2d(i, t) - 1)\}.$$

Maximize the logarithm of $\mathscr{L}$ with respect to $\tau(i)$ and β. Then for observation i for which $d(i, 1) = 1$ and $d(i, 2) = 1$, $\tau(i) \rightarrow \infty$. For observation i for which $d(i, 1) = 0$ and $d(i, 2) = 0$, $\tau(i) \rightarrow -\infty$. For the other observations the likelihood function may be concentrated in the $\tau(i)$ and then maximized with respect to β.

From the symmetry of the derivative of Φ it is obvious that $\tau(i) = -\beta/2$. (This is true for any symmetric Φ, not just the probit. Thus Andersen's analysis for the logit yields the same result.)

Substituting this expression into the likelihood function, it is straightforward to prove that if $\sigma_\tau^2 \rightarrow 0$, plim $\hat{\beta} = 2\beta$. If $\sigma_\tau^2 \rightarrow \infty$, plim $|\hat{\beta}| \rightarrow \infty$. The proof in the normal case for the first result closely parallels Andersen's proof. The second proof is trivial and hence is not given here.

Andersen (1973) demonstrates that conditioning the likelihood function on the event $\Sigma d(i, t) = 1$ eliminates the fixed effect in a logit model, so that conditional maximum likelihood estimators are consistent. (This result is originally due to Rasch 1960. See also the analysis of Haberman 1977.) This result is specific to a logit model. In a probit model such conditioning does not eliminate the fixed effect. In fact for the probit model the conditional maximum likelihood estimator generates unbounded estimators of the fixed effect for certain subsets of the observations.

Thus while conditional likelihood methods are helpful in a logit model, they are not helpful in a probit model, or in a general qualitative choice model. Conditioning the likelihood function in a logit model eliminates the fixed effect. First differences in a linear probability model eliminate the fixed effect, and thus the natural estimator for that model is linear regression. At this point the transformation of the probit function that eliminates the fixed effect is not known. The key point for the present discussion is that the lesson to extract from the binary logit case is *not* that conditional likelihood methods are a general approach in fixed effect models but that conditional likelihood methods in that case happen to be a convenient representation (or transformation) that eliminates fixed effects.

probit model with fixed effect) gives no guide to the performance of the fixed effect estimator when T is as large as 8, a feasible sample size in many data sets currently available. Recent Monte Carlo evidence by Wright and Douglas (1976) for the fixed effect Rasch-Andersen logit model finds that for T as small as 20 a fixed effect logit estimator performs as well as alternative consistent estimators.[10]

In Monte Carlo results reported in section 4.4 the fixed effect probit estimator performs well for T as small as 8 as long as no lagged values of dummy variables are included in the model. When lagged values are included, as is required to generate a first-order Markov model for discrete data, the fixed effect estimator performs badly.

In view of these results it is of interest to consider alternative solutions to the problem of initial conditions. An easily computed approximate solution to the problem of initial conditions is considered in the Monte Carlo experiments. This solution, which turns out to be relatively successful, approximates the reduced form marginal probability of the initial state in the sample by a probit function which has as its argument as much presample information on the exogenous variables as is available. The disturbance in the index variable that generates the reduced form probit function is left freely correlated with the structural errors over the sample period. This estimator works well in the Monte Carlo experiments, even though the probability of the initial state is not a probit function.

Specifically the following procedure is proposed and examined:

1. Approximate the probability of $d(i, J)$, the initial state in the sample, by a probit with index function

$$Y(i, J) = \sum_{l=0}^{J} D(l)[\mathbf{Z}(i,l)] + \mu(i,J)$$

$$Y(i,J) \geq 0 \text{ iff } d(i,J) = 1, \ Y(i,J) < 0 \quad \text{otherwise,}$$

where $D(l)[\mathbf{Z}(i, l)]$ is a general function of the $\mathbf{Z}(i, l)$, $l = 0, \ldots, J$, and $\mu(i, J)$ is assumed to be normally distributed with mean zero.[11]

2. Permit $\mu(i, J)$ to be freely correlated with $\varepsilon(i, t)$, $t = J + 1, \ldots, T$.

3. Estimate the model by the method of maximum likelihood without

[10] Their alternative consistent estimator is the conditional maximum likelihood estimator proposed by Rasch and Andersen and mentioned in the preceding note.

[11] In practice polynomials in $\mathbf{Z}(i, l)$ are used to form the $D(l)$ functions.

imposing any restrictions between the parameters of the structural system and the parameters of the approximate reduced form probability function for the initial state of the sample.

Assuming that the model is exact as $I \rightarrow \infty$, and for fixed T, the maximum likelihood estimator is consistent.

4.4 Some Monte Carlo Evidence

This section presents results from a limited set of Monte Carlo experiments. In each experiment with fixed effect estimators, 25 samples of 100 individuals ($I = 100$) are selected for 8 periods ($T = 8$). Results are presented for both a probit model with only exogenous explanatory variables and the first-order Markov model. The exogenous variables are assumed to follow a Nerlove process (Nerlove 1971):

$$Z(i,t) = 0.1t + 0.5Z(i,t-1) + \omega(i,t),$$

where $\omega(i, t)$ is a uniform random variable with mean zero and range $-1/2$ to $1/2$. This process well approximates the age-trended variables found in many microdata panel sets, especially in labor market analysis.

Results for the multivariate probit model with fixed effect but without lagged dummy variables are presented in table 4.1. Maximum likelihood fixed effect estimates are presented in the first three rows. The model generating the data is given at the bottom of the table. Samples are generated from a normal random number generator. The variance of $U(i, t)$ is set at one. The variance of the fixed effect, σ_τ^2, is changed for different experiments. For $\beta = 0.1$ the fixed effect estimator does well. The estimated value (denoted $\hat{\beta}$) comes very close to the true value. For $\beta = -1$, or $\beta = 1$, the estimator does not perform as well, but the bias is never more than 10 percent and is always toward zero. As the variance in the fixed effects decreases, so does the bias.[12]

These results are consistent with the findings of Wright and Douglas (1976) who use Monte Carlo methods to investigate the performance of the fixed effect logit estimator for the Rasch-Andersen model. In a study with panels of length $T = 20$ per person, they find that the fixed effect logit

[12] One unusual feature of these experiments is the consistent finding of a bias toward zero when a bias occurs. Andersen's two-period analysis would suggest an upward bias. The exogenous variables in the model investigated in the text have a much more complex character than the simple treatment effect variable used by Andersen.

estimator is virtually unbiased and its distribution is well described by a limiting normal distribution with variance-covariance matrix based on the estimated information matrix.

To judge the importance of the bias, one requires a benchmark. The benchmark selected in this chapter is a random effect estimator that integrates out the fixed effect. This is a multivariate probit model with random effect as presented in Heckman and Willis (1975) or chapter 3, section 3.5.

For each set of parameter values 25 samples with 100 observations of 3 periods are generated. The random effect estimator with $T = 3$ costs roughly the same to compute as the fixed effect estimator with $T = 8$. In terms of computational cost, the two estimators are equivalent.

The results with this model are presented in the final two rows of table 4.1. For a variance of $\sigma_\tau^2 = 3$, the random effect estimator displays more bias than the fixed effect estimator. For $\sigma_\tau^2 = 1$, the two estimators do about equally well. These experiments suggest that there is no clear ranking of the two estimators.

Test statistics for the random effect estimator (not given in the table) based on the estimated information matrix lead to rejection of the false null hypothesis that $\beta = 0$ far more often than test statistics based on the information matrix for the fixed effect estimator. On the basis of this limited evidence, if the estimators are to be used to make inference, the random effect estimator seems preferable.

Next consider some Monte Carlo experiments with the fixed effect estimator for a first-order Markov process. The results from these experiments are displayed in the first part of table 4.2. The same Nerlove process that generates the exogenous variables used in the preceding experiments is used in these experiments. The process operates for 25 periods before samples of 8 periods for each of the 100 individuals used in the 25 samples for each parameter set are selected.

The fixed effect probit estimator performs badly. The bias is greatest for large values of the variance in person effects (σ_τ^2) and when there are no exogenous variables in the model. But even the smallest bias reported in the table is still bad. The t statistics based on the estimated information matrix result in misleading inferences. From experimental results not reported in the table, one does not reject the false null hypothesis of $\gamma = \beta = 0$ in the vast majority of samples.

Note that estimates of γ are downward biased and estimates of β are upward biased. These results are very similar to Nerlove's Monte Carlo

Table 4.1
Monte Carlo results for models without lagged variables[a]

Values of $\hat{\beta}$ for the fixed effect probit model[b]

	$\beta = 1$	$\beta = -0.1$	$\beta = -1$
$\sigma_\tau^2 = 3$	0.90	-0.10	-0.94
$\sigma_\tau^2 = 1$	0.91	-0.09	-0.95
$\sigma_\tau^2 = 0.5$	0.93	-0.10	-0.96

Values of $\hat{\beta}$ for the random effect probit model[c]

	$\beta = 1$	$\beta = -1$
$\sigma_\tau^2 = 3$	1.15	-0.85
$\sigma_\tau^2 = 1$	1.04	-0.92

[a] The model generating the data is
$Y(i,t) = Z(i,t)\beta + \tau(i) + U(i,t)$,
$i = 1, \ldots, I$; if $Y(i,t) \geq 0$, $d(i,t) = 1$, $t = 1, \ldots, T$, otherwise, $d(i,t) = 0$.
$Z(i,t)$ is generated by the Nerlove (1971) process,
$Z(i,t) = 0.1t + 0.5Z(i,t-1) + \omega(i,t)$
$\omega(i,t) \sim U[-0.5, 0.5]$.
[b] $I = 100$, $T = 8$.
[c] $I = 100$, $T = 3$.

results in a linear equation model analogue of the Markov model (Nerlove 1971). Fixed effect estimators generate a downward-biased estimate of the coefficient of the lagged value of the endogenous variable in that model, just as they do for the state dependence coefficient in the Markov model.

In view of the poor performance of the fixed effect estimator as a solution to the problem of initial conditions, it is of some interest to examine the performance of some alternative estimators. The middle section of table 4.2 reports the results of a limited Monte Carlo study of the approximate random effect estimator proposed in section 4.3. The samples used to generate these estimates are the first three periods of the data utilized in the samples of the Monte Carlo study of the Markov model estimated by the fixed effect probit scheme. The first-period marginal probability is assumed to depend solely on first-period values of the exogenous variables. The proposed approximate random effect estimator discussed in section 4.3 does somewhat better than the fixed effect estimator. The $\hat{\gamma}$ consistently overstates the true γ and $\hat{\beta}$ understates the true β. As σ_{τ}^2 declines, so does the bias in the estimator. In results not reported here, t statistics are much more reliable in this model than in the fixed effect probit model since they lead to correct inference in a greater proportion of the samples.[13]

As in the discussion of the fixed effect probit model with strictly exogenous explanatory variables, it is natural to seek a suitable benchmark with which to compare the performance of the proposed estimators. One benchmark that provides an ideal case is a model with known nonstochastic initial conditions. Twenty-five samples with 100 observations of 3 periods are generated for each set of parameter values. Random effect maximum likelihood estimates for this model are presented in the final section of table 4.2. While this estimator is less biased than the approximate estimator, it is nonetheless biased. The difference between the results for the approximate random effect estimator for the case of stochastic initial conditions and the results for the estimator with known initial conditions

[13] Another ad hoc estimator was tried. This estimator fits a linear probability function to predict the marginal probability of the first sample period state. The predicted value is substituted in place of the actual value, and the γ parameter associated with the state in the first period is permitted to be distinct from the γ parameter for the sample transition. In terms of bias the performance of this estimator is intermediate between the fixed effect estimator and the proposed approximate random effect estimator given in the text, even when several years of presample data on the exogenous variables are used to predict the probability of the first sample period state. The estimator works well for the special problem of testing the null hypothesis of no state dependence ($\gamma = 0$).

Table 4.2
Monte Carlo results for models with lagged variables[a]

Values of $\hat{\gamma}$ and $\hat{\beta}$ for the fixed effects estimator[b]

		$\sigma_\tau^2 = 3$			$\sigma_\tau^2 = 1$		
		$\beta = -0.1$	$\beta = 1$	$\beta = 0$	$\beta = -0.1$	$\beta = 1$	$\beta = 0$
$\gamma = 0.5$	$\hat{\gamma}$	0.14	0.19	0.03	na[d]	0.25	0.17
	$\hat{\beta}$	−0.07	1.21	—	na[d]	1.17	—
$\gamma = 0.1$	$\hat{\gamma}$	−0.34	−0.21	−0.04	−0.28	−0.15	−0.01
	$\hat{\beta}$	−0.06	1.14	—	−0.08	1.12	—

Values of $\hat{\gamma}$ and $\hat{\beta}$ for the proposed approximate random effect estimation[c]

		$\sigma_\tau^2 = 3$			$\sigma_\tau^2 = 1$		
		$\beta = -0.1$	$\beta = 1$	$\beta = 0$	$\beta = -0.1$	$\beta = 1$	$\beta = 0$
$\gamma = 0.5$	$\hat{\gamma}$	0.63	0.60	0.70	na[d]	0.54	0.62
	$\hat{\beta}$	−0.131	0.91	—	na[d]	0.93	—
$\gamma = 0.1$	$\hat{\gamma}$	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 random effect estimator with known nonstochastic initial conditions[c]

		$\sigma_\tau^2 = 3$		
		$\beta = -0.1$	$\beta = 1$	$\beta = 0$
$\gamma = 0.5$	$\hat{\gamma}$	na[d]	0.57	na[d]
	$\hat{\beta}$	na[d]	0.94	—
$\gamma = 0.1$	$\hat{\gamma}$	0.13	0.12	0.14
	$\hat{\beta}$	−0.11	1.10	—

[a] The model generating the data is

$$Y(i,t) = Z(i,t)\beta + \gamma d(i,t-1) + \tau(i) + U(i,t)$$
$$Y(i,t) \geq 0 \Leftrightarrow d(i,t) = 1 \quad \text{and} \quad Y(i,t) = 0 \Leftrightarrow d(i,t) = 0.$$

The process operates 25 periods before samples are generated. $Z(i,t)$ is generated by the Nerlove proces (see table 4.1).

[b] $I = 100$, $T = 8$.

[c] $I = 100$, $T = 3$.

[d] Data are not available because the model was not estimated.

indicates that the approximate estimator is not badly biased relative to an ideal alternative.

There is one disquieting feature in the results presented in table 4.2. All of the maximum likelihood estimators exhibit considerable bias. This point deserves much further examination in view of the increasingly widespread use of maximum likelihood methods for the analysis of discrete data models.

4.5 Conclusions

This chapter has examined the problem of initial conditions that arises in estimating a discrete time-discrete data stochastic process when serially correlated unobservable variables generate the process. The analysis has been confined to a first-order Markov process, although the issues discussed here apply to the general class of models with structural state dependence considered in chapter 3. Exact and approximate solutions are proposed, and Monte Carlo evidence is presented on the performance of certain simple estimators likely to be used in applied work.

One solution based on a fixed effect probit model is simple to compute and hence is attractive. However, a limited set of Monte Carlo experiments reveals that the fixed effect probit model performs badly in panels of length eight when lagged dummy variables are included as explanatory variables, as is required to generate the general stochastic process proposed in chapter 3. When the explanatory variables of the model are all strictly exogenous, the fixed effect probit estimator performs well.

An alternative approximate simple estimator is proposed and examined, and this estimator is found to perform well in a limited set of Monte Carlo experiments. A disquieting finding in most of the experiments is that for panels with 100 observations of 3 periods, which are large samples by current standards, the method of maximum likelihood produces biased estimators even under ideal conditions. This finding deserves much further study.

More than the usual cautionary note is required to qualify the results of the Monte Carlo experiments reported in this chapter. Only a limited set of experiments has been performed. Accordingly the Monte Carlo evidence reported here is best thought of as suggestive rather than definitive. Much further work is required and would be very desirable.

References

Amemiya, T. 1978. A Note on the Estimation of a Time Dependent Markov Chain Model. Stanford University.

Andersen, E. B. 1973. *Conditional Inference and Models for Measuring*. Copenhagen: Mentalhygiejnisk Forsknings Institut.

Anderson, T. W. 1976. Panels and Time Series Analysis: Markov Chains and Autoregressive Processes. Technical report 24. Department of Statistics, Stanford University.

Anderson, T. W., and L. Goodman. 1957. Statistical Inference about Markov Chains. *Annals of Mathematical Statistics*. 28: 89–110.

Haberman, S. 1977. Maximum Likelihood Estimates in Exponential Response Models. *The Annals of Statistics*. 5: 815–841.

Heckman, J., and R. Willis. 1975. Estimation of a Stochastic Model of Reproduction: An Econometric Approach. In *Household Production and Consumption*, ed. N. Terleckyj. National Bureau of Economic Research, New York.

Karlin, S., and H. Taylor. 1975. *A First Course in Stochastic Processes*. 2nd ed. New York: Academic Press.

Kiefer, J., and J. Wolfowitz. 1956. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Annals of Mathematical Statistics*. 27: 887–906.

Mundlak, Y. 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica*. 46: 69–86.

Nerlove, M. 1971. Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections. *Econometrica*. 39: 359–382.

Neyman, J., and E. Scott. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*. 16: 1–32.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute of Educational Research.

Wright, B. D., and G. Douglas. 1976. Better Procedures for Sample-Free Item Analysis. Research memorandum 20. Statistical Laboratory, Department of Education, University of Chicago.

III STRUCTURAL DISCRETE PROBABILITY MODELS DERIVED FROM THEORIES OF CHOICE