Binary choice panel data models with predetermined variables

Manuel Arellano[a,*], Raquel Carrasco[b]

[a] *CEMFI, Casado del Alisal 5, 28014 Madrid, Spain*
[b] *Departamento de Economía, Universidad Carlos III de Madrid, Madrid, Spain*

Abstract

We present a class of binary choice models for panel data with the following features: (i) The explanatory variables are predetermined but not strictly exogenous. This includes lagged dependent variables as well as other forms of unspecified feedback. (ii) Individual effects are allowed to be correlated with the explanatory variables. Dependence is specified through the conditional expectation of the effects which is let to be non-parametric. We also present a GMM estimator for these models, which is consistent and asymptotically normal for fixed T and large N. We study its finite sample properties in an autoregressive model by means of Monte Carlo simulations. Finally, as an empirical illustration, we estimate a female labour force participation equation with predetermined children using PSID data.

JEL classification: C23

Keywords: Discrete choice; Panel data; Predetermined variables; Random effects; Labor force participation

1. Introduction

It is well known that parameter estimates from short panels jointly estimated with the individual effects can be seriously biased when the explanatory variables are only predetermined as opposed to strictly exogenous. This situation includes models with lagged dependent variables as well as other models in which the explanatory variables are Granger-caused by the endogenous variables. In linear models with additive effects the standard response to this problem has been to consider instrumental-variables estimates that exploit the lack of correlation between future errors in first-differences

* Corresponding author. Tel.: +34-914-290-551; fax: +34-914-291-056.
E-mail address: arellano@cemfi.es (M. Arellano).

and lagged values of the variables (e.g. Anderson and Hsiao, 1981; Holtz-Eakin et al., 1988, or Arellano and Bond, 1991). However, much fewer results are available on discrete choice models with predetermined variables and other non-linear models of interest in microeconometrics.

The purpose of this paper is to develop a class of semi-parametric random effects binary choice models without the strict exogeneity assumption. Random effects models with only strictly exogenous variables have been considered by Chamberlain (1980, 1984) and Newey (1994a). Heckman (1981a,b) studied discrete choice models with state dependence and random effects. A different strand of the literature, beginning with the conditional logit formulation of Andersen (1970), has considered "fixed effects" specifications in which the full distribution of the effects is left unrestricted (or treated as "non-parametric"). This includes the maximum score method proposed by Manski (1987), which relaxes the logit assumption but requires strict exogeneity and stationarity, and the models considered by Honoré and Kyriazidou (2000), and Honoré and Lewbel (2002). Honoré and Kyriazidou include a lagged dependent variable, but their remaining explanatory variables are also required to be strictly exogenous. Honoré and Lewbel allow for other predetermined variables but require the presence of a continuous, strictly exogenous, explanatory variable that is independent of the effects. For a survey of the literature, see Arellano and Honoré (2001).

Fixed effects can be regarded as a random effects specification that leaves the distribution of the effects unrestricted. They are attractive as a way to ensure that the conditional distribution of the effects does not play a role in identifying the parameters of interest. However, sometimes one may be willing to impose a certain amount of structure in the dependence between the effects and the endogenous variables if in exchange this makes it possible to relax other aspects of the economic model of interest. In such situations, a semi-parametric random effects specification may represent a useful compromise. In this regard, the semi-parametric random effects models considered in this paper contain a non-parametric conditional expectation of the effects given the predetermined variables, but are otherwise parametric.

An example where lack of strict exogeneity would be expected even after controlling for individual effects, is given by the effect of children in female labour force participation decisions. In such context, assuming that children are strictly exogenous is much stronger than the assumption of predeterminedness, since it would require us to maintain that labour supply plans have no effect on fertility decisions at any point in the life cycle (Browning, 1992, p. 1462). Here feedback effects from lagged participation decisions (or lagged shocks to participation) to current and future children outcomes cannot be ruled out. The result is that the identification arrangements and the estimation techniques that are useful with strictly exogenous variables break down.

The paper is organized as follows. Section 2 presents the model and a GMM estimator for this model, which is consistent and asymptotically normal for fixed T as N goes to infinity. In Section 3 we study the finite sample properties of this estimator in a binary choice model with a single lagged dependent variable by means of Monte Carlo simulations. Finally, in Section 4, as an empirical illustration, we estimate a female labour force participation equation with predetermined children and individual effects using PSID data.

2. Models and estimators

2.1. The model

Let us consider the following error-components binary choice model for N individuals observed T consecutive time periods

$$y_{it} = \mathbf{1}(\gamma_t + \beta x_{it} + u_{it} \geqslant 0) \quad (i = 1,\ldots,N;\ t = 1,\ldots,T), \tag{2.1}$$

$$u_{it} = \eta_i + v_{it}, \tag{2.2}$$

where $\mathbf{1}(.)$ denotes the indicator function, so that y_{it} is a $0-1$ variable. T is small and N is large. Let us also denote $w_{it} = (x_{it}, y_{i(t-1)})$ and $w_i^t = (w_{i1} \cdots w_{it})$.[1] The composite errors u_{it} are assumed to have a known distribution up to scale given w_i^t, for example normal, of the form:

$$u_{it} \mid w_i^t \sim \mathcal{N}(\mathrm{E}(\eta_i \mid w_i^t),\ \sigma_t^2). \tag{2.3}$$

The sequence of conditional means $\{\mathrm{E}(\eta_i \mid w_i^s),\ s=1,\ldots,T\}$ is left unrestricted. They are just linked by the law of iterated expectations:

$$\mathrm{E}(\eta_i \mid w_i^t) = \mathrm{E}(\mathrm{E}(\eta_i \mid w_i^{t+1}) \mid w_i^t). \tag{2.4}$$

Thus, the model allows for dependence between the explanatory variable x_{it} and the individual effect η_i through the conditional mean of the latter given the observed time path of w. Moreover, the model specifies x as a predetermined variable, in the sense that while x_{it} does not depend on current or future values of the transitory error v_{it}, there may be feedback from lagged values of v or y to x_{it}. Assumption (2.3), however, essentially excludes serially correlated errors. This is so because the mean independence condition $\mathrm{E}(v_{it} \mid w_i^t) = 0$, while not implying by itself serial mean independence of the v_{it}, it does rule out standard patterns of autocorrelation.

The conditional probabilities specified by the model are

$$\Pr(y_{it} = 1 \mid w_i^t) = \Phi\left(\frac{\gamma_t + \beta x_{it} + \mathrm{E}(\eta_i \mid w_i^t)}{\sigma_t}\right), \tag{2.5}$$

where $\Phi(.)$ is the standard normal cdf. The assumption of normality is unessential and could be replaced by any other parametric assumption. Notice that since the model is conditional on w_i^t it could include $y_{i(t-1)}$ as an additional regressor, and indeed the next section focuses on the case where $x_{it} = y_{i(t-1)}$. However, the specification above was chosen for its simplicity, and also to emphasize the fact that, in order to allow for x_{it} to depend upon lagged y, we ought to condition on the histories of both x and y, even in the absence of an independent effect of $y_{i(t-1)}$ on y_{it} given x_{it}, η_i and v_{it}.

The model could also accommodate a situation where feedback effects are present for some explanatory variables but not others. Effectively, if there is a strictly exogenous x, all its lags and leads will be included in the conditioning set at each t.

[1] The variable y_{i0} is assumed to be observed to simplify the notation.

It is of some interest to relate the present model to a model in which (2.3) is replaced by the assumption

$$v_{it} \mid w_i^t, \eta_i \sim \mathcal{N}(0, \omega_t^2), \tag{2.6}$$

so that

$$\Pr(y_{it} = 1 \mid w_i^t, \eta_i) = \Phi\left(\frac{\gamma_t + \beta x_{it} + \eta_i}{\omega_t}\right). \tag{2.7}$$

Note that if for *some* t

$$\eta_i \mid w_i^t \sim \mathcal{N}(\mathrm{E}(\eta_i \mid w_i^t), \sigma_{\eta t}^2), \tag{2.8}$$

it can be easily shown that this assumption together with (2.6) implies an expression identical to (2.5) with $\sigma_t^2 = \sigma_{\eta t}^2 + \omega_t^2$. It would thus appear that assumptions (2.3) and (2.6) are connected through (2.8). However, if $\eta_i \mid w_i^t$ is normal, since w_i^t contains y_i^{t-1} which is a sequence of binary variables, it follows that $\eta_i \mid w_i^{t-1}$ cannot in general be normal, and therefore an expression of the form of (2.5) could not hold for $\Pr(y_{i(t-1)} = 1 \mid w_i^{t-1})$. So, the two models are not nested. Unlike in (2.6), in model (2.1)–(2.3) η_i and v_{it} are not assumed to be conditionally independent, and in general they will be correlated.[2]

The present model's identification rests on the assumption that the demeaned error $u_{it} - \mathrm{E}(\eta_i \mid w_i^t)$ has a distribution that may change with t, but is independent of the individual's history w_i^t. Since the history will affect the shape of the conditional distributions $\eta_i \mid w_i^t$, our assumption implies that in general v_{it} will only be mean independent of w_i^t, which is a limitation of this approach.

Thus, a feature of our model is that it matters to the distributional assumption if one starts observing the individuals one period earlier or later. That is, if $u_{it} \mid w_{i1} \cdots w_{it}$ is normally distributed, in general $u_{it} \mid w_{i2} \cdots w_{it}$ will be distributed as a normal mixture and hence non-normal. This is an undesirable mathematical property since typically in applications individuals are not necessarily observed from the date in which the process started.

Notice that this is also true of the static random effects probit model of Chamberlain (1984) and Newey (1994a). The assumption in such case is

$$\eta_i + v_{it} \mid x_i^T \sim \mathcal{N}(\mathrm{E}(\eta_i \mid x_i^T), \sigma_t^2).$$

If the assumption is intended to hold for any T, it follows that $v_{it} \mid x_i^T$ may be normal for some T but not for *any* T (except in very special cases).

We motivated assumption (2.3) as a distributional specification for the random effects model represented by (2.1) and (2.2). Nevertheless, taken together, assumptions (2.1)–(2.3) can be given an alternative interpretation, which suggests a different class of applications of the techniques developed below. Namely, suppose that the model of economic interest is

$$y_{it} = \mathbf{1}(\gamma_t + \beta x_{it} + \mathrm{E}(\eta_i \mid w_i^t) + \varepsilon_{it} \geqslant 0),$$

[2] Notice that unless v_{it} is a purely expectational error (an innovation), there will normally be reasons for assuming correlation between η and v_t (e.g. random preferences, omitted time-varying characteristics, etc.).

where the forecast $\mathrm{E}(\eta_i \mid w_i^t)$ is part of the structural equation. The latent variable η_i is unobserved to the agent, and its forecast is revised each period as information accumulates. Such model together with the distributional assumption for the error ε_{it}

$$\varepsilon_{it} \mid w_i^t \sim \mathcal{N}(0, \sigma_t^2)$$

is equivalent to (2.1)–(2.3).

2.2. *Discrete predetermined variables*

2.2.1. *Identification*

We start by considering identification and estimation in the case where x_{it} is a discrete random variable with a finite support of J points. It is useful to consider this case since the model becomes fully parametric, while effectively leaving the distribution of x_{it} unrestricted. The continuous case is taken up below.

The vector w_{it} will have a finite support of $2J$ points given by $(\phi_1 \cdots \phi_{2J})$. The $t \times 1$ random vector x_i^t has a multinomial distribution, and takes J^t different values. Similarly, the vector w_i^t takes on $(2J)^t$ different values ϕ_j^t $(j = 1, \ldots, (2J)^t)$.

As a matter of notational convenience we order the ϕ_j^t such that for $t > 1$:

$$\phi_j^t = (\phi_j^{t-1}, \phi_\ell) \tag{2.9}$$

with $j = (\ell - 1)(2J)^{t-1} + 1, \ldots, \ell(2J)^{t-1}$; $\ell = 1, \ldots, (2J)$. That is, for a specific value $w_i^{t-1} = \phi_j^{t-1}$ there are $2J$ different values of $w_i^t = (w_i^{t-1}, w_{it})$ with $w_i^{t-1} = \phi_j^{t-1}$, which we represent as $w_i^t = \phi_j^t$ ordered as in (2.9).

Let us denote

$$p_{tj} = \Pr(y_{it} = 1 \mid w_i^t = \phi_j^t) \equiv h_t(\phi_j^t) \quad (j = 1, \ldots, (2J)^t) \tag{2.10}$$

and

$$\psi_j^t = \mathrm{E}(\eta_i \mid w_i^t = \phi_j^t) \quad (j = 1, \ldots, (2J)^t). \tag{2.11}$$

Therefore we have

$$p_{tj} = \Phi\left(\frac{\gamma_t + \beta\phi_j^{[t]} + \psi_j^t}{\sigma_t}\right), \tag{2.12}$$

where $\phi_j^{[t]}$ denotes the last element of the vector ϕ_j^t. By the law of iterated expectations we also have

$$\psi_j^{t-1} = \sum_{\ell=1}^{2J} \psi_{(\ell-1)(2J)^{t-1}+j}^t \Pr(w_{it} = \phi_\ell \mid w_i^{t-1} = \phi_j^{t-1}) \times (j = 1, \ldots, (2J)^{(t-1)};\ t = 2, \ldots, T). \tag{2.13}$$

Moreover, since the model includes a constant term, it is not restrictive to assume that $\mathrm{E}(\eta_i) = 0$. Therefore

$$\mathrm{E}(\eta_i) = \sum_{\ell=1}^{2J} \mathrm{E}(\eta_i \mid w_{i1} = \phi_\ell) \Pr(w_{i1} = \phi_\ell) = 0. \tag{2.14}$$

Let us consider the partition $\phi_j = (\phi_{1j}, \phi_{2j})$ where ϕ_{2j} is either 0 or 1. Then the probabilities in (2.13) factorize as

$$\begin{aligned}\Pr(w_{it} = \phi_\ell \mid w_i^{t-1} = \phi_j^{t-1}) = \Pr(x_{it} = \phi_{1\ell} \mid w_i^{t-1} = \phi_j^{t-1}, y_{i(t-1)} = \phi_{2\ell}) \\ \times \Pr(y_{i(t-1)} = \phi_{2\ell} \mid w_i^{t-1} = \phi_j^{t-1}).\end{aligned} \tag{2.15}$$

Notice that the second term on the right-hand side contains the probabilities specified by the model. The first term consists of unspecified conditional probabilities for the x, and so they are additional reduced form parameters:

$$\begin{aligned}\pi_{t\ell}^{jk} &= \Pr(x_{it} = \phi_{1\ell} \mid w_i^{t-1} = \phi_j^{t-1}, y_{i(t-1)} = \phi_{2k}) \\ &(t = 2, \ldots, T;\ \ell = 1, \ldots, J;\ j = 1, \ldots, (2J)^{t-1};\ k = 0, 1).\end{aligned} \tag{2.16}$$

The probabilities $p_\ell = \Pr(w_{i1} = \phi_\ell)$ are also left unrestricted and just add $2J$ parameters to the full likelihood function of the data.

The number of reduced form parameters p_{tj} is $(2J + (2J)^2 + \cdots + (2J)^T)$, and the number of $\pi_{t\ell}^{jk}$ is $2(2J + (2J)^2 + \cdots + (2J)^{T-1})$.

The coefficients can be estimated up to scale. Using σ_1 as the scale, we can estimate γ_t/σ_1, β/σ_1, ψ_j^t/σ_1 and the relative scales σ_1/σ_t. We shall use $\sigma_1 = 1$ as the normalization restriction.

The structural parameters are $\beta, \gamma_1 \ldots \gamma_T$, $\sigma_2 \ldots \sigma_T$ and the ψ_j^t. The number of parameters ψ_j^t is $((2J) + \cdots + (2J)^T)$, although they are subject to $\sum_{j=0}^{T-1} (2J)^j$ restrictions.[3] In conclusion, the total number of estimating equations is

$$r = 4 \sum_{j=1}^{T-1} (2J)^j + (2J)^T + (2J) + 1 \tag{2.17}$$

while the number of parameters to be estimated including the $\pi_{t\ell}^{jk}$ and the initial probabilities is

$$k = 3 \sum_{j=1}^{T-1} (2J)^j + (2J)^T + (2J) + 2T. \tag{2.18}$$

Hence the number of overidentifying restrictions is

$$r - k = \sum_{j=1}^{T-1} (2J)^j - 2T + 1. \tag{2.19}$$

[3] We could alternatively say that the required parameters are

$$\psi_j^T = \mathrm{E}(\eta_i \mid w_i^T = \phi_j^T) \qquad (j = 1, \ldots, (2J)^T)$$

since the remaining ψ_j^t are functions of those and the cell probabilities.

Identification of β up to scale requires that at least $T \geqslant 2$, or that there are at least three observations available on each individual (since we assumed that y_{i0} is observed). With two observations, contrary to the situation in the linear model, β would only be identified under homoskedasticity. Indeed, setting $J = 2$, $\sigma_1 = \sigma_2 = 1$, γ_t constant, and $\Pr(y_{i0} = 1) = 1$, a straightforward calculation reveals that

$$\beta = \frac{\sum_{\ell=0}^{1} \sum_{j=0}^{1} \Phi^{-1}(p_{\ell j})\pi_{\ell j} p_j - \Phi^{-1}(p_1)}{\pi_{11} p_1 + \pi_{10} p_0 - 1}, \tag{2.20}$$

where

$$p_j = \Pr(y_{i1} = j \,|\, x_{i1} = 1) \quad (j = 0, 1),$$

$$p_{\ell j} = \Pr(y_{i2} = 1 \,|\, x_{i2} = \ell,\ x_{i1} = 1,\ y_{i1} = j),$$

$$\pi_{\ell j} = \Pr(x_{i2} = \ell \,|\, x_{i1} = 1,\ y_{i1} = j) \quad (\ell = 0, 1;\ j = 0, 1).$$

In the derivation of (2.20), we are assuming that x_{it} is a $0 - 1$ variable, and that there are effectively two observations on each unit since $y_{i0} = 1$ with probability one. The coefficient β is nevertheless identified due to the homoskedasticity assumption $\sigma_1 = \sigma_2$.[4]

2.2.2. *Minimum distance and maximum likelihood estimation*

Let us define the variables

$$d_{ij}^t = \mathbf{1}(w_i^t = \phi_j^t). \tag{2.21}$$

Then the unrestricted maximum likelihood estimator of p_{tj} is given by

$$\hat{p}_{tj} = \frac{1}{\sum_{i=1}^{N} d_{ij}^t} \sum_{i=1}^{N} y_{it} d_{ij}^t \quad (t = 1, \ldots T;\ j = 1, \ldots, (2J)^t). \tag{2.22}$$

Similarly, for $\pi_{t\ell}^{jk}$ we have

$$\hat{\pi}_{t\ell}^{jk} = \frac{1}{\sum_{i=1}^{N} d_{ij}^{t-1} \mathbf{1}(y_{i(t-1)} = k)} \sum_{i=1}^{N} \mathbf{1}(x_{it} = \phi_{1\ell}) d_{ij}^{t-1} \mathbf{1}(y_{i(t-1)} = k) \tag{2.23}$$

and for the initial probabilities:

$$\hat{p}_\ell = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(w_{i1} = \phi_\ell) \quad (\ell = 1, \ldots, 2J). \tag{2.24}$$

[4] We obtained the expression for β in (2.20) as the solution to the moment equation

$$\mathrm{E}\{x_1[\mathrm{E}(\eta \,|\, x_2, x_1, y_1) - \mathrm{E}(\eta \,|\, x_1)]\} = 0,$$

where $\mathrm{E}(\eta \,|\, x_2, x_1, y_1) - \mathrm{E}(\eta \,|\, x_1) = \Phi^{-1}[\Pr(y_2 = 1 \,|\, x_2, x_1, y_1)] - \Phi^{-1}[\Pr(y_1 = \,|\, x_1)] - \beta \Delta x_2$. This type of moment will be used below for GMM estimation.

We can form the vector[5]

$$g(\hat{p},\theta) = \begin{pmatrix} \sum_{\ell=1}^{2J} \psi_\ell^1 \hat{p}_\ell \\ \hat{p}_{tj} - \Phi\left(\dfrac{\gamma_t + \beta\phi_j^{[t]} + \psi_j^t}{\sigma_t}\right) \\ \psi_j^{t-1} - \sum_{\ell=1}^{J} \psi^t_{(\ell-1)(2J)^{t-1}+j} \hat{\pi}^{j1}_{t\ell}\, \hat{p}_{(t-1)j} - \sum_{\ell=J+1}^{2J} \psi^t_{(\ell-1)(2J)^{t-1}+j} \hat{\pi}^{j0}_{t(\ell-J)}(1-\hat{p}_{(t-1)j}) \\ \vdots \end{pmatrix}. \tag{2.25}$$

The vector of functions $g(\hat{p},\theta)$ includes the terms for all j and t. The vector $\hat{p}$ contains the $\hat{p}_{tj}, \hat{\pi}^{jk}_{t\ell}$ and $\hat{p}_\ell$, while θ contains all the parameters to be estimated. A minimum distance (MD) estimator of θ solves

$$\hat{\theta} = \arg\min_\theta g(\hat{p},\theta)' A_N g(\hat{p},\theta), \tag{2.26}$$

where A_N is a consistent estimate of the inverse of the covariance matrix of $g(\hat{p},\theta)$.

As an alternative to the MD procedure, the model can be estimated by maximum likelihood. The contribution to the log-likelihood for the ith observation is given by

$$\begin{aligned} L_i = &\sum_{t=1}^{T} [y_{it} \ln p_{it} + (1-y_{it}) \ln(1-p_{it})] \\ &+ \sum_{t=2}^{T} \left\{ \sum_{\ell=1}^{J} \mathbf{1}(x_{it} = \phi_{1\ell}) \sum_{j=1}^{(2J)^{t-1}} d_{ij}^{t-1} [y_{i(t-1)} \ln \pi^{j1}_{t\ell} + (1-y_{i(t-1)}) \ln \pi^{j0}_{t\ell}] \right\} \\ &+ \sum_{\ell=1}^{2J} \mathbf{1}(w_{i1} = \phi_\ell) \ln p_\ell, \end{aligned} \tag{2.27}$$

where

$$p_{it} = \Phi\left(\frac{\gamma_t + \beta x_{it} + \sum_{j=1}^{(2J)^t} \psi_j^t d_{ij}^t}{\sigma_t}\right) \tag{2.28}$$

and the ψ_j^t are solved recursively using the restrictions (2.13) and (2.14) as functions of ψ_j^T and the other parameters of the model. The log-likelihood is maximized as a function of the $\gamma_t, \sigma_t, \beta, \psi_j^T, \pi^{jk}_{t\ell}$ and p_ℓ.

[5] We have implicitly ordered the observations in such a way that $\phi_{2\ell} = 1$ for $\ell = 1,\ldots,J$ and $\phi_{2\ell} = 0$ for $\ell = J+1,\ldots,2J$.

If N is not sufficiently large relative to the number of cells—which depends on J and T—some cells may contain very few or no observations. Those cells will be trimmed, thereby reducing the effective number of first-stage parameters and equations available for MD estimation. The finite sample properties of the MD estimates may be affected by the degree of first-stage trimming. The problem of choosing the amount of trimming in this context, therefore, resembles the choice of the number of instruments in instrumental variable models.

2.2.3. GMM estimation

The following simpler method avoids the joint estimation of the parameters of interest with the nuisance coefficients ψ_j^t. By inverting Eq. (2.5) we obtain

$$\sigma_t \Phi^{-1}[h_t(w_i^t)] = \gamma_t + \beta x_{it} + \mathrm{E}(\eta_i \mid w_i^t). \tag{2.29}$$

First-differencing this equation we have

$$\sigma_t \Phi^{-1}[h_t(w_i^t)] - \sigma_{t-1}\Phi^{-1}[h_{t-1}(w_i^{t-1})] - \Delta\gamma_t - \beta\Delta x_{it} = \varepsilon_{it} \tag{2.30}$$

where $\Delta\gamma_t = \gamma_t - \gamma_{t-1}$, $\Delta x_{it} = x_{it} - x_{i(t-1)}$, and

$$\varepsilon_{it} = \mathrm{E}(\eta_i \mid w_i^t) - \mathrm{E}(\eta_i \mid w_i^{t-1}). \tag{2.31}$$

Therefore

$$\mathrm{E}(\varepsilon_{it} \mid w_i^{t-1}) = 0. \tag{2.32}$$

Notice that

$$h_t(w_i^t) = \sum_{j=1}^{(2J)^t} d_{ij}^t p_{tj}. \tag{2.33}$$

Moreover, the conditional moment restriction (2.32) is equivalent to the following unconditional moments (see Chamberlain, 1987, p. 308):

$$\mathrm{E}(d_{ij}^{t-1}\varepsilon_{it}) = 0 \quad (j = 1, \ldots, (2J)^{t-1}) \tag{2.34}$$

or:

$$\mathrm{E}\left\{ d_{ij}^{t-1}\left[\sigma_t\Phi^{-1}\left(\sum_{j=1}^{(2J)^t} d_{ij}^t p_{tj}\right) - \sigma_{t-1}\Phi^{-1}\left(\sum_{j=1}^{(2J)^{t-1}} d_{ij}^{t-1} p_{(t-1)j}\right) - \Delta\gamma_t - \beta\Delta x_{it}\right]\right\} = 0. \tag{2.35}$$

The orthogonality conditions corresponding to the p_{tj} are

$$\mathrm{E}[d_{ij}^t(y_{it} - p_{tj})] = 0 \quad (j = 1, \ldots, (2J)^t), \tag{2.36}$$

$$\mathrm{E}[d_{ij}^{t-1}(y_{i(t-1)} - p_{(t-1)j})] = 0 \quad (j = 1, \ldots, (2J)^{t-1}). \tag{2.37}$$

The complete set of moment conditions can be used to obtain joint estimates of the p_{tj} and the coefficients of interest. However, since the latter are unrestricted moments

there is no efficiency loss (as far as the estimation of the parameters of interest is concerned) in replacing in the first set of orthogonality conditions (2.35) unrestricted estimates of the p_{tj} and the $p_{(t-1)j}$.

Letting z_{it} be a vector containing the indicators $d_{ij}^{t-1}(j = 1, \ldots, (2J)^{t-1})$, and

$$\hat{h}_t(w_i^t) = \sum_{j=1}^{(2J)^t} d_{ij}^t \hat{p}_{tj} \tag{2.38}$$

a two-step GMM method can be based on the sample orthogonality conditions

$$\frac{1}{N}\sum_{i=1}^{N} z_{it}(\sigma_t \Phi^{-1}[\hat{h}_t(w_i^t)] - \sigma_{t-1}\Phi^{-1}[\hat{h}_{t-1}(w_i^{t-1})] - \Delta\gamma_t - \beta\Delta x_{it}) \times (t = 2, \ldots, T) \tag{2.39}$$

yielding asymptotically efficient estimates of β, $\Delta\gamma_t$ and σ_t subject to the normalization restriction $\sigma_1 = 1$. Since both y_i^T and x_i^T have finite supports the model is fully parametric and the asymptotic distribution of the estimators can be obtained using standard GMM asymptotic theory.

An optimal weighting matrix is given by a consistent estimate of the inverse asymptotic covariance matrix of the orthogonality conditions (2.39). Notice that since the latter depends on the joint limiting distribution of $z_{it}\varepsilon_{it}$ and $\hat{p}_{tj}$, a standard outer-product formula using the estimated moments would be inappropriate.

In practice, the number of available moment conditions may be substantially smaller than $\sum_t (2J)^{t-1}$ since z_{it} will only contain the indicators corresponding to outcomes that occurred in the data. Moreover, the sample moments will only depend on estimated h_t of non-empty cells. Detailed illustrations of these issues are provided below using Monte Carlo simulations and an empirical application.[6]

Minimum distance estimation of binary choice models using inverted probabilities was first proposed by Berkson (1944) for data with many observations per cell (see Amemiya, 1985, p. 275). Transformation (2.30) is also similar to the one employed by Newey (1994a) for a probit model with only strictly exogenous variables. In the strictly exogenous case, however, the error term ε_{it} does not appear since there is no sequential updating of the conditional expectations of the individual effects.

2.3. *Continuous predetermined variables*

If x_{it} is a continuous random variable, estimation cannot be based on cell sample frequencies. Instead we now rely on non-parametric smoothed estimators of the reduced form probabilities $h_t(w_i^t)$ in order to construct orthogonality conditions. Another aspect

[6] We expect the large N, fixed T distribution of the GMM estimator to lack a bias term even if x is discrete with infinite support (Delgado and Mora, 1995). On account of finite sample performance, however, it may be desirable to use a smoothed estimator of the probabilities, or to drop cells that contain very few observations (as we do in the empirical application).

is that with a continuous x_{it} it is feasible to estimate non-parametrically the distributions of the composite errors for each t, although this will not be pursued here. [7]

The $t \times 1$ random vector $y_i^{t-1} = (y_{i0}, \ldots, y_{i(t-1)})$ still has a multivariate Bernouilli distribution, and takes on 2^t different values ζ_j^{t-1} $(j = 1, \ldots, 2^t)$. Therefore, we consider non-parametric estimates of $h_t(w_i^t)$ of the form

$$\tilde{h}_t(w_i^t) = \sum_{j=1}^{2^t} \tilde{h}_{tj}(x_i^t) \mathbf{1}(y_i^{t-1} = \zeta_j^{t-1}), \tag{2.40}$$

where $\tilde{h}_{tj}(x_i^t)$ is a non-parametric smooth (e.g. kernel) estimator of the conditional probability

$$h_{tj}(x_i^t) = \Pr(y_{it} = 1 \mid x_i^t, y_i^{t-1} = \zeta_j^{t-1}). \tag{2.41}$$

Contrary to the multinomial case, now the conditional moment restrictions given by (2.32) do not imply a finite number of orthogonality conditions. Here we do not consider the issue of selecting and estimating optimal instruments (which would be required for asymptotic efficiency), and merely exploit the moment conditions $\mathrm{E}(w_i^{t-1} \varepsilon_{it}) = 0$. Let us define

$$\tilde{\psi}_{it}(\theta) = \begin{pmatrix} 1 \\ w_i^{t-1} \end{pmatrix} \{\sigma_t \Phi^{-1}[\tilde{h}_t(w_i^t)] - \sigma_{t-1} \Phi^{-1} \times [\tilde{h}_{t-1}(w_i^{t-1})] - \Delta\gamma_t - \beta \Delta x_{it}\} \tag{2.42}$$

where $\theta = (\beta, \Delta\gamma_2, \ldots, \Delta\gamma_T, \sigma_2, \ldots, \sigma_T)$. Let the sample orthogonality conditions be given by

$$\tilde{b}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} (\tilde{\psi}_{i2}(\theta)', \ldots, \tilde{\psi}_{iT}(\theta)')'. \tag{2.43}$$

A semi-parametric two-step GMM estimator of θ solves

$$\tilde{\theta} = \arg\min_{\theta} \tilde{b}_N(\theta)' A_N \tilde{b}_N(\theta) \tag{2.44}$$

where A_N is a weighting matrix.

To illustrate why the previous moment conditions may be expected to satisfy the rank condition and hence be sufficient to identify the parameters, let us consider the case where $T = 2$. In such case the 3×1 vector $\theta = (\beta, \Delta\gamma_2, \sigma_2)$ would be just identified provided the following submatrix of the derivatives of the moment conditions has full rank:

$$\Psi = \begin{pmatrix} \mathrm{E}(x_{i1} \Delta x_{i2}) & -\mathrm{E}(x_{i1} \Phi^{-1}[h_2(w_i^2)]) \\ \mathrm{E}(y_{i0} \Delta x_{i2}) & -\mathrm{E}(y_{i0} \Phi^{-1}[h_2(w_i^2)]) \end{pmatrix},$$

[7] Chen (1998) generalized the model of Newey (1994a) with strictly exogenous variables by allowing the distribution of the composite errors to be unknown.

which under the model's assumptions takes the form

$$\Psi = \begin{pmatrix} \mathrm{E}(x_{i1}\Delta x_{i2}) & -\mathrm{E}(x_{i1}[\gamma_2 + \beta x_{i2} + \mathrm{E}(\eta_i \mid w_i^2)]) \\ \mathrm{E}(y_{i0}\Delta x_{i2}) & -\mathrm{E}(y_{i0}[\gamma_2 + \beta x_{i2} + \mathrm{E}(\eta_i \mid w_i^2)]) \end{pmatrix}.$$

Note that in a model without individual effects and $\beta = 0$, the rank condition will be satisfied unless $\mathrm{E}(y_{i0}) = 0$.[8]

As long as the order and rank conditions are satisfied, estimation could be based on a smaller set of moments using as instruments, for example, a subset of the lagged variables contained in w_i^{t-1}. Nevertheless, the form of the error ε_{it} in (2.30) will remain unchanged since the relevant non-parametric probabilities correspond to those specified by the model for the full vector of conditioning variables.

Under appropriate regularity conditions (see Newey, 1994a; Newey and McFadden, 1994):

$$\sqrt{N}\tilde{b}_N(\theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_0) \tag{2.45}$$

with

$$V_0 = \mathrm{E}[(\psi_i(\theta) + a_i)(\psi_i(\theta) + a_i)'], \tag{2.46}$$

where $\psi_i(\theta) = (\psi_{i2}(\theta)' \cdots \psi_{iT}(\theta)')'$, $\psi_{it}(\theta) = w_i^{t-1}\varepsilon_{it}$, and a_i is an adjustment term that takes into account the fact that the $h_t(w_i^t)$ have been replaced by non-parametric estimates. There may be a need for trimming in the first-stage non-parametric estimation, in which case the second-stage moments $\tilde{b}_N(\theta)$ will be based on a smaller effective sample size.

Following Newey (1994b), V_0 can be consistently estimated by mean of

$$\tilde{V} = \frac{1}{N}\sum_{i=1}^{N}(\tilde{\psi}_i + \tilde{a}_i)(\tilde{\psi}_i + \tilde{a}_i)', \tag{2.47}$$

where $\tilde{\psi}_i = \psi_i(\tilde{\theta})$ and

$$\tilde{a}_i = \sum_{s=2}^{T}\sum_{j=1}^{2^s}\tilde{a}_{sji} \tag{2.48}$$

with

$$\tilde{a}_{sji} = \frac{1}{\sum_{k=1}^{N}\mathbf{1}(y_k^{s-1} = \phi_j^{s-1})}\sum_{k=1}^{N}\left(\frac{\partial\tilde{\psi}_k}{\partial h_{sj}}\right) y_{is}K_{sj}(x_k^s - x_i^s)\mathbf{1}(y_k^{s-1} = \phi_j^{s-1}) \tag{2.49}$$

and $K_{sj}(.)$ is the kernel used in the estimation of $h_{sj}(x_i^s)$.

The advantage of (2.49) is that it does not require the calculation of a_i as an explicit functional of the parameters, the data, and the non-parametric probabilities. This expression, however, can be obtained in our case as a direct application of Proposition 5

[8] Our moment conditions are linear in the parameters given the non-parametric components, so that the Jacobian Ψ is constant. Here we just consider the dependence of Ψ on the true values of the parameters.

in Newey (1994a), which provides the basis for an alternative estimator of a_i. We illustrate the result when $T = 2$ and the following moment is used:

$$\psi_i(\theta) = \bar{w}_i^1 \{\sigma_2 \Phi^{-1}[h_2(w_i^2)] - \Phi^{-1}[h_1(w_i^1)] - \Delta\gamma_2 - \beta\Delta x_2\},$$

where $\bar{w}_i^1 = (1, y_{i0}, x_{i1})$ and $\theta = (\beta, \Delta\gamma_2, \sigma_2)$. In this case $a_i = a_{1i} + a_{2i}$ and

$$a_{1i} = -\bar{w}_i^1 \phi\{\Phi^{-1}[h_1(w_i^1)]\}^{-1}[y_{i1} - h_1(w_i^1)],$$

$$a_{2i} = -\bar{w}_i^1 \sigma_2 \phi\{\Phi^{-1}[h_2(w_i^2)]\}^{-1}[y_{i2} - h_2(w_i^2)].$$

Finally, a consistent estimate of the asymptotic variance matrix of $\tilde{\theta}$ is given by

$$(\tilde{D}' A_N \tilde{D})^{-1} \tilde{D}' A_N \tilde{V} A_N \tilde{D} (\tilde{D}' A_N \tilde{D})^{-1}, \tag{2.50}$$

where

$$\tilde{D} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \tilde{\psi}_i}{\partial \theta'}. \tag{2.51}$$

Alternatively, one could use a bootstrap estimator of the asymptotic variance matrix of $\tilde{\theta}$.

2.4. *Marginal effects of interest*

The effect of changing x_{it} from x' to x'' on the probability of $y_{it} = 1$ is given by

$$\begin{aligned} \Delta_t^S(x', x'') &= \Pr(\gamma_t + \beta x'' + u_{it} \geqslant 0) - \Pr(\gamma_t + \beta x' + u_{it} \geqslant 0) \\ &= G_t(\gamma_t + \beta x'') - G_t(\gamma_t + \beta x'), \end{aligned} \tag{2.52}$$

where $G_t(.)$ is the marginal *cdf* of $-u_{it}$. According to our specification G_t can be written as

$$G_t(r) = \mathrm{E}\left[\Phi\left(\frac{r + \mathrm{E}(\eta_i \mid w_i^t)}{\sigma_t}\right)\right]. \tag{2.53}$$

To obtain a consistent estimate of $\Delta_t^S(x', x'')$, a consistent estimate of $\sigma_t^{-1}\mathrm{E}(\eta_i \mid w_i^t)$ is required. In view of our estimation strategy and (2.29), a natural choice is

$$\widehat{\sigma_t^{-1}}\mathrm{E}(\eta_i \mid w_i^t) = \Phi^{-1}[\hat{h}_t(w_i^t)] - \hat{\sigma}_t^{-1}(\hat{\gamma}_t + \hat{\beta} x_{it}). \tag{2.54}$$

Upon substitution, the estimated effect is[9]

$$\begin{aligned} &\hat{\Delta}_t^S(x', x'') \\ &= \frac{1}{N} \sum_{i=1}^{N} \{\Phi(\hat{\sigma}_t^{-1}\hat{\beta}(x'' - x_{it}) + \Phi^{-1}[\hat{h}_t(w_i^t)]) \\ &\quad - \Phi(\hat{\sigma}_t^{-1}\hat{\beta}(x' - x_{it}) + \Phi^{-1}[\hat{h}_t(w_i^t)])\}. \end{aligned} \tag{2.55}$$

[9] A related discussion is contained in Chamberlain (1984, pp. 1272–1274).

Note that, although β is constant, the estimated effect is period-specific as a result of conditioning sequentially over time.

It is of some interest to distinguish $\Delta_t^S(x',x'')$ from the *predictive effect* of changing x_{it} from x' to x'' given by

$$\Delta_t^P(x',x'') = \mathrm{E}\{\Pr(y_{it}=1 \mid y_i^{t-1},x_i^{t-1},x_{it}=x'') - \Pr(y_{it}=1 \mid y_i^{t-1},x_i^{t-1},x_{it}=x')\}. \quad (2.56)$$

Estimation of $\Delta_t^P(x',x'')$ does not require the use of the model. It is a reduced form effect that can be estimated using the non-parametric estimates of the sequential conditional probabilities $\hat{h}_t(w_i^t)$. Note that $\Delta_t^S(x',x'')$ gives the structural effect of x_t on y_t in the observed population of u_t, while $\Delta_t^P(x',x'')$ mixes the direct effect of x_t with the indirect effect due to the dependence between η and x_t.

In a model with two (or more) continuous explanatory variables, direct relative marginal effects can also be obtained. These are constant over time and given by the ratios of the corresponding β coefficients. For example, we have

$$\frac{\beta_1}{\beta_2} = \frac{\partial G_t(\gamma_t + \beta_1 x_{1it} + \beta_2 x_{2it})/\partial x_{1it}}{\partial G_t(\gamma_t + \beta_1 x_{1it} + \beta_2 x_{2it})/\partial x_{2it}}. \quad (2.57)$$

2.5. Individual effects interacted with time effects

A simple extension of the basic framework outlined above is a model in which individual effects are interacted with time effects given by

$$y_{it} = \mathbf{1}(\gamma_t + \beta x_{it} + \eta_i \delta_t + v_{it} \geqslant 0) \quad (2.58)$$

and

$$\eta_i \delta_t + v_{it} \mid w_i^t \sim \mathcal{N}(\mathrm{E}(\eta_i \mid w_i^t)\delta_t, \sigma_t^2). \quad (2.59)$$

In this model, estimation can be based on the transformation

$$\sigma_t \Phi^{-1}[h_t(w_i^t)] - r_t \sigma_{t-1} \Phi^{-1}[h_{t-1}(w_i^{t-1})] = (\gamma_t - r_t \gamma_{t-1}) + \beta x_{it} - r_t \beta x_{i(t-1)} + \varepsilon_{it}^*, \quad (2.60)$$

where $r_t = \delta_t/\delta_{(t-1)}$ and $\mathrm{E}(\varepsilon_{it}^* \mid w_i^{t-1}) = 0$. A normalization restriction such as $\delta_1 = 1$ is required.

3. Experimental evidence

3.1. State dependence with unobserved heterogeneity

In this section we study the finite sample properties of the ML and GMM estimators described above in a binary choice model with a single lagged dependent variable by means of Monte Carlo simulations. The model is given by

$$y_{it} = \mathbf{1}(\gamma + \alpha y_{i(t-1)} + \eta_i + v_{it} \geqslant 0) \quad (t = 2, \ldots, T). \quad (3.1)$$

This is a model of independent interest, whose basic motivation is to facilitate the distinction between unobserved heterogeneity and state dependence in the analysis of binary-state discrete-time processes. One example is the analysis of sequences of employment and unemployment states, where a substantive question is whether or not unemployment causes future unemployment (e.g. Heckman (1981c) or Card and Sullivan (1988), who use these models for measuring the effect of training programs on employment and unemployment probabilities). Another example is the analysis of a housing quality indicator over time as in the work by Moon and Stotsky (1993). Moon and Stotsky consider the effect of rent control on a two state housing condition variable (sound and unsound) allowing for state dependence and unobserved heterogeneity.

We assume that the composite error given $y_i^{t-1} = (y_{i1}, \ldots, y_{i(t-1)})$ has a logistic distribution of the form:

$$\eta_i + v_{it} \mid y_i^{t-1} \sim \text{Logistic}(\text{E}(\eta_i \mid y_i^{t-1}), \sigma^2). \tag{3.2}$$

Since we do not expect significant differences between the performance of logit and probit models, we chose the logistic function F, say, because the inverse probabilities have an explicit form and their calculation is faster. [10]

The form of the likelihood for one individual, conditional on the first observation, is therefore $\mathscr{L}_i = \prod_{t=2}^{T} (F_{it})^{y_{it}} (1 - F_{it})^{(1-y_{it})}$ where

$$F_{it} = F\left(\gamma + \alpha y_{i(t-1)} + \sum_{j=1}^{2^{(t-1)}} \psi_j^{t-1} d_{ij}^{t-1}\right),$$

$$\psi_j^{t-1} = \text{E}(\eta_i \mid y_i^{t-1} = \phi_j^{t-1}) \tag{3.3}$$

and

$$d_{ij}^{t-1} = \mathbf{1}(y_i^{t-1} = \phi_j^{t-1}) \quad (j = 1, \ldots, 2^{t-1}).$$

The coefficients ψ_j^{t-1} are solved recursively using $\text{E}(\eta_i) = 0$ and the law of iterated expectations, as functions of ψ_j^{T-1} and the other parameters of the model. The likelihood function is particularly simple in this case since the conditioning variables are binary, and only the probabilities specified by the model are required in the evaluation of the ψ_j^{t-1}. We further simplified the specification by keeping γ and σ^2 constant over time, so that the coefficients should be interpreted as being relative to σ.

The distribution of the initial observation $p_1 = \Pr(y_{i1} = 1)$ is left unrestricted in our model. However, since p_1 enters the restriction $\text{E}(\eta_i) = \psi_1^1 p_1 + \psi_2^1 (1 - p_1) = 0$, we maximized the full likelihood of the data given by

$$\mathscr{L}_i \times (p_1)^{y_{i1}} (1 - p_1)^{(1-y_{i1})}.$$

[10] The probit transformation Φ^{-1} does not have an explicit form, but it can be easily evaluated numerically (see, for example, Beasley and Springer, 1977).

3.2. Some alternative likelihoods

In the simulations we also considered an alternative likelihood model conditional on an individual effect with a mass point distribution. Assuming that the conditional random variables $\eta_i \mid y_{i1} = 1$ and $\eta_i \mid y_{i1} = 0$ are discrete with finite support given by m mass points $e_1, \ldots, e_m$, the likelihood for one individual given the initial observation in this case is

$$\mathscr{L}_{\mathscr{H}i} = \sum_{\ell=1}^{m} \prod_{t=2}^{T} G_{it}(e_\ell)^{y_{it}} [1 - G_{it}(e_\ell)]^{(1-y_{it})} \Pr(\eta_i = e_\ell \mid y_{i1} = \phi_1), \tag{3.4}$$

where

$$G_{it}(e_\ell) = F(\gamma + \alpha y_{i(t-1)} + e_\ell). \tag{3.5}$$

$\mathscr{L}_{\mathscr{H}i}$ is a function of γ, α, the mass points $e_1, \ldots, e_m$ and the conditional probabilities $\Pr(\eta_i = e_\ell \mid y_{i1} = \phi_1)$.

Notice that the likelihood (3.4) is based on a decomposition of the joint distribution of y_i^T given by

$$p(y_i^T) = p(y_{i1}) \int \prod_{t=2}^{T} p(y_{it} \mid y_i^{t-1}, \eta_i) \, \mathrm{d}H(\eta_i \mid y_{i1}) \tag{3.6}$$

which should be distinguished from the following alternative decomposition:

$$p(y_i^T) = \int \prod_{t=2}^{T} p(y_{it} \mid y_i^{t-1}, \eta_i) \, p(y_{i1} \mid \eta_i) \, \mathrm{d}H(\eta_i). \tag{3.7}$$

By specifying the conditional distributions of η_i given y_{i1}, $H(\eta_i \mid y_{i1})$, in (3.6) as opposed to the marginal distribution of η_i, $H(\eta_i)$ (as done in (3.7)), we allow for dependence between y_{i1} and η_i, while leaving the initial conditions of the process unrestricted. Decomposition (3.6) is akin to the estimators of linear autoregressive models that leave the initial conditions unrestricted. If on the other hand one wishes to specify the distribution $p(y_{i1} \mid \eta_i)$ (for example, by assuming some form of stationarity), (3.7) would be the relevant decomposition.

Specifying a mass point distribution for η_i is attractive because it is flexible, and also because by letting the support of η grow with sample size it is often possible to establish asymptotic properties for the estimators with respect to a model with an unrestricted distribution for η (cf. Heckman and Singer, 1984).

The likelihood $\mathscr{L}_{\mathscr{H}i}$ only entails estimating two conditional distributions of the effects in the first-order autoregressive case. However, in the more general model with predetermined variables considered in the previous section, the number of conditional distributions would be larger. Moreover, the conditional distribution of the predetermined variables given the unobserved component would be required to be able to construct the mixing likelihood. This would cause the predetermined variables to become fully endogenous, since we would effectively need a specification of the joint

distribution of y and x given η.[11] As a consequence, this approach may not be feasible even with discrete explanatory variables, and even less so with continuous variables.

In the simulations we specified two mass points $\{e_1, e_2\}$ for the random variables $\eta_i \mid y_{i1}$. As a result, the likelihood function contained five free parameters. Namely, γ, α, $\Pr(\eta_i = e_1 \mid y_{i1} = 1)$, $\Pr(\eta_i = e_1 \mid y_{i1} = 0)$, and e_1. Notice that given one mass point and the associated probabilities, the other mass point is determined by the condition $\mathrm{E}(\eta_i) = 0$.

Another approach that we considered is Chamberlain (1985)'s conditional autoregressive logit estimator for the model

$$\Pr(y_{it} = 1 \mid y_i^{t-1}, \eta_i) = F(\alpha y_{i(t-1)} + \eta_i),$$

(see also Magnac, 2000). When $T = 4$, this method is based on the observation that $\Pr(y_{i2} = 1 \mid y_{i4}, y_{i2} + y_{i3} = 1, y_{i1}, \eta_i)$ is independent of η_i. The corresponding conditional MLE of α turns out to be of the form

$$\tilde{\alpha}_{CML} = \ln(n_1/n_2),$$

where n_1 and n_2 are the number of observations with $(1,1,0,0)$ or $(0,0,1,1)$ in the first case, and $(1,0,1,0)$ or $(0,1,0,1)$ in the second. An expression along the same lines can be obtained for larger values of T. However, extensions to more general models are problematic. In a model with only strictly exogenous variables in addition to lagged y, it leads to estimators with a slower than root-N rate of convergence (Honoré and Kyriazidou, 2000), and little is known about models with other predetermined variables.

In addition, given that one of the original motivations was the bias from estimators that attempt to estimate the fixed effects in short panels, we decided to include a logit ML estimator that does this.

Finally, as a benchmark we calculated maximum likelihood estimates conditional on y_{i1} without unobserved heterogeneity. The likelihood is given by

$$\mathscr{L}_{\mathscr{R}i} = \prod_{t=2}^{T} (F_{rit})^{y_{it}} (1 - F_{rit})^{(1-y_{it})}, \tag{3.8}$$

where

$$F_{rit} = F(\gamma + \alpha y_{i(t-1)}). \tag{3.9}$$

The function $\mathscr{L}_{\mathscr{R}i}$ is a special case of $\mathscr{L}_i$ and the other likelihoods. We generated data without unobserved heterogeneity, so that the maximizers of the five likelihood functions are consistent estimates of the same parameters (a description of the design

[11] A discussion of this problem in the context of a duration model with predetermined time-varying covariates can be found in Bover et al. (2002). The form of a likelihood comparable to (3.4) when an additional predetermined discrete variable x_{it} is present is

$$\sum_{\ell=1}^{m} \prod_{t=2}^{T} G_{it}(e_\ell)^{y_{it}} [1 - G_{it}(e_\ell)]^{(1-y_{it})} \Pr(x_{it} \mid x_i^{t-1}, y_i^{t-1}, \eta_i = e_\ell) \Pr(\eta_i = e_\ell \mid y_{i1}, x_{i1}).$$

of the experiments is given below). Thus, we hope to assess the efficiency loss incurred in allowing for the various forms of unobserved heterogeneity relative to the homogeneous model. [12]

3.3. GMM estimation

We now turn to describe the GMM estimator of model (3.2) used in the Monte Carlo analysis. By using moment conditions of the type given in (2.39), only the coefficient α would be estimated. However, γ is also a parameter of interest since the dynamics of the process (3.1) is determined by both γ and α. The parameter γ is identifiable from the orthogonality conditions

$$\mathrm{E}(\mathrm{E}(\eta_i \mid y_i^{t-1})) = \mathrm{E}(F^{-1}[\Pr(y_{it} = 1 \mid y_i^{t-1})] - \gamma - \alpha y_{i(t-1)}) = 0$$
$$(t = 2, \ldots, T). \tag{3.10}$$

We therefore obtained joint estimates of γ and α relying on both first-difference and levels sample orthogonality conditions as follows:

$$b_{1tN} = \frac{1}{N}\sum_{i=1}^{N} z_{it}(F^{-1}[\hat{h}_t(y_i^{t-1})] - F^{-1}[\hat{h}_{t-1}(y_i^{t-2})] - \alpha \Delta y_{i(t-1)})$$
$$(t = 3, \ldots, T), \tag{3.11}$$

$$b_{2tN} = \frac{1}{N}\sum_{i=1}^{N} (F^{-1}[\hat{h}_t(y_i^{t-1})] - \gamma - \alpha y_{i(t-1)}) \quad (t = 2, \ldots, T), \tag{3.12}$$

where z_{it} is a subset of the indicators d_{ij}^{t-2} $(j = 1, \ldots, 2^{t-2})$ corresponding to outcomes that occurred in the data, and the $\hat{h}_t(y_i^{t-1})$ denote the sample frequencies

$$\hat{h}_t(y_i^{t-1}) = \hat{p}_{tj} \equiv \frac{1}{\sum_{k=1}^{N} d_{kj}^{t-1}} \sum_{k=1}^{N} y_{kt} d_{kj}^{t-1} \quad \text{if } y_i^{t-1} = \phi_j^{t-1} \quad (j = 1, \ldots, 2^{t-1}).$$

Note that the sample moments will only depend on relative frequencies of non-empty cells (those with $\sum_{k=1}^{N} d_{kj}^{t-1} > 0$), so that depending on the values of N and the model's parameters the actual number of $\hat{p}_{tj}$'s involved in estimation may be substantially smaller than $\sum_t 2^{t-1}$.

The GMM results reported below are for "one-step" estimators given by

$$(\hat{\gamma}, \hat{\alpha}) = \arg\min \sum_{t=3}^{T} b_{1tN}' \left(\sum_{i=1}^{N} z_{it} z_{it}'\right)^{-1} b_{1tN} + \sum_{t=2}^{T} b_{2tN}' b_{2tN}. \tag{3.13}$$

The estimates in (3.13) use a non-optimal but convenient weighting matrix which does not require the calculation of preliminary consistent estimates of γ and α.

[12] It would also be of interest to consider a Monte Carlo design with heterogeneity generated according to one of the previous models. However, in such case we would only have pseudo-true parameters for the remaining models, since they are not nested. In Section 4 we report a simulation with heterogeneity in the context of the application.

In the calculation of the orthogonality conditions b_{1tN} and b_{2tN}, we used the following modification of the logit transformation mentioned by Cox (1970) and Amemiya (1985, p. 278):

$$F^{-1}(\hat{p}_{tj}) = \ln\left(\frac{\hat{p}_{tj} + (2n_{tj})^{-1}}{1 - \hat{p}_{tj} + (2n_{tj})^{-1}}\right), \tag{3.14}$$

where n_{tj} is the number of observations in the cell with $y_i^{t-1} = \phi_j^{t-1}$, so that $n_{tj} = \sum_{i=1}^{N} d_{ij}^{t-1}$. This modification has the advantage that the transformation is still defined if $\hat{p}_{tj} = 0$ or 1. Moreover, its mean bias relative to $\ln[p_{tj}/(1 - p_{tj})]$ can be shown to be of a smaller order of magnitude than the bias for the standard logit transformation.

3.4. Experimental design

We generated longitudinal observations from a homogeneous stationary first-order Markov process. Thus, relative to model (3.1), in the data generating process we have $\eta_i = 0$ with probability one. Moreover

$$p_{10} = \Pr(y_{it} = 1 \mid y_{i(t-1)} = 0) = F(\gamma), \tag{3.15}$$

$$p_{11} = \Pr(y_{it} = 1 \mid y_{i(t-1)} = 1) = F(\gamma + \alpha). \tag{3.16}$$

The degree of dependence in the process can be measured by

$$\rho = corr(y_{it}, y_{i(t-1)}) = p_{11} - p_{10} = F(\gamma + \alpha) - F(\gamma) \tag{3.17}$$

while the stationary probability is given by

$$p^* = \Pr(y_{it} = 1) = \frac{p_{10}}{1 - (p_{11} - p_{10})}. \tag{3.18}$$

So it seemed natural to start by setting combinations of values for ρ and p^*, from which the implied values of γ and α can be derived using the logistic transformation.

Although ρ and p^* are natural descriptive quantities for our data generating process, we are mainly concerned with the sampling distributions of estimates of α and γ. The reason is that the coefficients in the linear index, like α and γ, will be typically parameters of interest in econometric applications in which the index is related to the agents' objective functions evaluations. Moreover, we would expect the coefficients in the linear index to remain well-defined with heterogeneous and non-stationary data.

We considered cases with $\rho = 0.2,\ 0.5$ and $p^* = 0.2,\ 0.5,\ 0.8$, which produced the following values for γ and α

	$\rho = 0.2$			$\rho = 0.5$		
p^*	0.2	0.5	0.8	0.2	0.5	0.8
γ	−1.66	−0.4	0.57	−2.2	−1.1	−0.4
α	1.08	0.81	1.08	2.6	2.2	2.6

One would expect that the larger the value of ρ the more difficult it becomes to distinguish between state dependence and unobserved heterogeneity. So large values of ρ

may produce quite imprecise estimates of γ and α. As an indication of some empirically relevant quantities, for the PSID sample that we use in Section 4, the female labour force participation rate is 0.55, while the gross first-order autocorrelation in participation is 0.65, (which reflects the combined effect of state dependence and heterogeneity). The estimates reported in Section 4, allowing for unobserved heterogeneity, imply a value of ρ of 0.31.

For each experiment, we generated 100 samples with $N = 500, 1000$ and $T = 4, 6$. With $T = 4$ the GMM estimates were based on 9 moment conditions (6 in first differences and 3 in levels) and 14 cell frequencies, whereas with $T = 6$ they were based on 35 moments (30 in differences and 5 in levels) and 62 cell frequencies.

3.5. *Monte Carlo results*

Tables 1–4 report means, percentage bias, standard deviations and root mean squared errors (MSE) for the GMM and ML estimates of model (3.1)–(3.2). The tables also report results for the alternative likelihood model with mass point distributions for the effects given y_{i1} (labelled ML-MP), for the conditional autoregressive logit model (in the case of α, labelled CML), and for the homogeneous model (labelled RML). Table 5 contains the results from the ML method that estimates α jointly with the fixed effects for the subset of experiments with $N = 500$ and $\rho = 0.2$.

The results for the experiments with $T = 4$ are contained in Tables 1 and 2. The comparison between GMM and ML in those tables shows that GMM almost always has a higher MSE for both γ and α than ML. ML tends to have a smaller standard deviation and bias than GMM. However, the differences between the two estimators are small except in the less favorable cases. The bias in the GMM estimate of γ is worryingly large when $\rho = 0.5$ and $p^* = 0.8$. More generally, it is noticeable that for the larger value of ρ the estimates of γ and α are less precise and have higher MSE than for those with $\rho = 0.2$. Also, for a given ρ, large or small values of $p*$ tend to produce worse estimates of α.

Turning to the comparison between GMM/ML with the homogeneous ML (RML), it turns out that the standard deviation of RML is between 1.5 and 3 times smaller than that of GMM or ML, with the difference becoming wider in the least favourable cases. RML is the estimator with the smallest bias and standard deviation. This result is to be expected since RML does not allow for unobserved heterogeneity, which is in fact absent from the data.

As far as ML-MP is concerned, the estimator of γ exhibits a greater MSE than those of GMM and ML, while in the case of α the result is the opposite. ML-MP estimates of γ are seriously downward biased when $\rho = 0.5$ and $p^* = 0.8$, even with $N = 1000$. However, ML-MP estimates of α have substantially smaller standard deviations than GMM/ML estimates. The poor performance of ML-MP estimates of γ relative to those of α is probably due to larger correlations between the estimates of γ and the estimates of the mass points. The ML-MP results may be sensitive to the number of mass points allowed in the conditional distributions of η, but we did not explore this issue.

With $T = 6$, the GMM estimates always have a smaller MSE than with $T = 4$, but this is due to reductions in variance that offset larger biases in all the experiments.

Table 1
Means and standard deviations of the estimators, $N = 500$, $T = 4$

	GMM	ML	ML-MP	RML	GMM	ML	ML-MP	CML	RML
($\beta = 0.2$, $p^* = 0.2$)		$\gamma = -1.66$					$\alpha = 1.08$		
Mean	−1.65	−1.68	−1.79	−1.66	0.99	1.10	0.97	1.12	1.06
Mean bias (%)	0.6	1.2	7.8	0.0	8.3	1.8	10.2	3.7	1.8
St. dev.	0.09	0.09	0.35	0.07	0.30	0.29	0.17	0.34	0.14
Root MSE	0.09	0.09	0.37	0.07	0.31	0.29	0.20	0.34	0.14
($\beta = 0.2$, $p^* = 0.5$)		$\gamma = -0.4$					$\alpha = 0.81$		
Mean	−0.38	−0.43	−0.38	−0.39	0.76	0.85	0.73	0.80	0.78
Mean bias (%)	5.0	7.5	5.0	2.5	6.2	4.9	9.9	1.2	3.7
St. dev.	0.13	0.10	0.16	0.07	0.22	0.18	0.14	0.21	0.11
Root MSE	0.13	0.11	0.16	0.07	0.23	0.18	0.16	0.21	0.11
($\beta = 0.2$, $p^* = 0.8$)		$\gamma = 0.57$					$\alpha = 1.08$		
Mean	0.60	0.54	0.74	0.57	1.03	1.13	0.98	1.08	1.06
Mean bias (%)	5.3	5.3	29.8	0.0	4.6	4.6	9.2	0.0	1.8
St. dev.	0.26	0.25	0.35	0.11	0.31	0.30	0.19	0.31	0.13
Root MSE	0.26	0.25	0.39	0.11	0.32	0.30	0.22	0.31	0.13
($\beta = 0.5$, $p^* = 0.2$)		$\gamma = -2.2$					$\alpha = 2.6$		
Mean	−2.16	−2.23	−2.32	−2.20	2.39	2.71	2.48	3.18	2.58
Mean bias (%)	1.8	1.4	5.4	0.0	8.1	4.2	4.6	22.3	0.8
St. dev.	0.13	0.13	0.36	0.09	0.51	0.47	0.23	2.47	0.16
Root MSE	0.13	0.13	0.38	0.09	0.55	0.48	0.26	2.54	0.16
($\beta = 0.5$, $p^* = 0.5$)		$\gamma = -1.1$					$\alpha = 2.2$		
Mean	−1.06	−1.11	−1.07	−1.08	2.12	2.23	2.13	2.21	2.16
Mean bias (%)	3.6	0.9	2.7	1.8	3.6	1.4	3.2	0.4	1.8
St. dev.	0.17	0.18	0.18	0.07	0.32	0.33	0.16	0.37	0.11
Root MSE	0.17	0.18	0.18	0.07	0.33	0.33	0.17	0.37	0.12
($\beta = 0.5$, $p^* = 0.8$)		$\gamma = -0.4$					$\alpha = 2.6$		
Mean	−0.26	−0.47	−0.19	−0.38	2.43	2.70	2.48	3.08	2.58
Mean bias (%)	35.0	17.5	52.5	5.0	6.5	3.8	4.6	18.5	0.8
St. dev.	0.44	0.34	0.46	0.11	0.53	0.43	0.23	2.09	0.13
Root MSE	0.46	0.35	0.51	0.11	0.56	0.44	0.26	2.14	0.13

100 replications. Root MSE denotes root mean squared error.

In contrast with GMM, the finite sample biases of the ML estimates do not increase with T for a fixed N. Since biases of this type have been shown to be sensitive to the choice of weighting matrix in other contexts, it would be interesting to explore to what extent they can be removed by using an optimal weighting matrix.

CML leaves both initial conditions and the distribution of the effects unrestricted. So it is the most robust method, but at the expense of relying exclusively on a fraction of the observations. In terms of MSE, it tends to lie between GMM and ML, although typically with a smaller bias and a larger standard deviation. There is, however, some

Table 2
Means and standard deviations of the estimators, $N = 1000$, $T = 4$

	GMM	ML	ML-MP	RML	GMM	ML	ML-MP	CML	RML
($\beta = 0.2$, $p^* = 0.2$)		$\gamma = -1.66$					$\alpha = 1.08$		
Mean	−1.65	−1.66	−1.75	−1.66	1.03	1.08	1.00	1.09	1.07
Mean bias (%)	0.6	0.0	5.4	0.0	4.6	0.0	7.4	0.9	0.9
St. dev.	0.06	0.06	0.26	0.05	0.23	0.20	0.14	0.23	0.1
Root MSE	0.06	0.06	0.28	0.05	0.24	0.20	0.16	0.23	0.10
($\beta = 0.2$, $p^* = 0.5$)		$\gamma = -0.4$					$\alpha = 0.81$		
Mean	−0.40	−0.41	−0.40	−0.40	0.81	0.82	0.78	0.82	0.81
Mean bias (%)	0.0	2.5	0.0	0.0	0.0	1.2	3.7	1.2	0.0
St. dev.	0.08	0.08	0.12	0.05	0.15	0.14	0.09	0.14	0.07
Root MSE	0.08	0.08	0.12	0.05	0.15	0.14	0.10	0.14	0.07
($\beta = 0.2$, $p^* = 0.8$)		$\gamma = 0.57$					$\alpha = 1.08$		
Mean	0.58	0.56	0.66	0.57	1.06	1.10	1.03	1.08	1.07
Mean bias (%)	1.7	1.7	15.8	0.0	1.8	1.8	4.6	0.0	0.9
St. dev.	0.17	0.16	0.24	0.09	0.22	0.20	0.14	0.22	0.11
Root MSE	0.17	0.16	0.26	0.09	0.22	0.20	0.15	0.22	0.11
($\beta = 0.5$, $p^* = 0.2$)		$\gamma = -2.2$					$\alpha = 2.6$		
Mean	−2.19	−2.20	−2.29	−2.20	2.54	2.61	2.53	2.70	2.60
Mean bias (%)	0.4	0.0	4.1	0.0	2.3	0.38	2.7	3.8	0.0
St. dev.	0.09	0.08	0.28	0.06	0.34	0.31	0.17	0.42	0.11
Root MSE	0.09	0.08	0.29	0.06	0.35	0.31	0.18	0.43	0.11
($\beta = 0.5$, $p^* = 0.5$)		$\gamma = -1.1$					$\alpha = 2.2$		
Mean	−1.07	−1.11	−1.10	−1.09	2.14	2.20	2.13	2.21	2.18
Mean bias (%)	2.7	0.91	0.0	0.9	2.7	0.0	3.2	0.4	0.9
St. dev.	0.12	0.12	0.17	0.06	0.23	0.22	0.12	0.25	0.09
Root MSE	0.12	0.12	0.17	0.06	0.24	0.22	0.14	0.25	0.09
($\beta = 0.5$, $p^* = 0.8$)		$\gamma = -0.4$					$\alpha = 2.6$		
Mean	−0.33	−0.41	−0.27	−0.39	2.5	2.63	2.52	2.63	2.59
Mean bias (%)	17.5	2.5	32.5	2.5	3.8	1.1	3.1	1.1	0.4
St. dev.	0.24	0.30	0.34	0.09	0.30	0.37	0.19	0.51	0.12
Root MSE	0.25	0.30	0.36	0.09	0.32	0.37	0.21	0.51	0.12

100 replications. Root MSE denotes root mean squared error.

evidence of a higher probability of outliers in the CML, which is reflected in very high sample standard deviations for some of the experiments.

Table 5 reports ML estimates of α for $\rho = 0.2$ and $N = 500$, obtained jointly with those of the fixed effects. Due to computing limitations, results for higher values of N and ρ are not reported. The results in Table 5, however, indicate that the biases are in all cases very large indeed for both $T = 4$ and $T = 6$.

In conclusion, the Monte Carlo results for the GMM and ML estimates of our model suggest a similar pattern to that typically encountered in linear autoregressive models.

Table 3
Means and standard deviations of the estimators, $N = 500, T = 6$

	GMM	ML	ML-MP	RML	GMM	ML	ML-MP	CML	RML
($\beta = 0.2$, $p^* = 0.2$)		$\gamma = -1.66$					$\alpha = 1.08$		
Mean	−1.62	−1.71	−1.71	−1.66	0.90	1.17	1.00	1.05	1.06
Mean bias (%)	2.4	3.0	3.0	0.0	16.7	8.3	7.4	2.8	1.8
St. dev.	0.07	0.07	0.16	0.06	0.21	0.19	0.14	0.16	0.12
Root MSE	0.08	0.08	0.17	0.06	0.28	0.21	0.16	0.16	0.12
($\beta = 0.2$, $p^* = 0.5$)		$\gamma = -0.4$					$\alpha = 0.81$		
Mean	−0.35	−0.39	−0.39	−0.39	0.70	0.80	0.76	0.79	0.79
Mean bias (%)	12.5	2.5	2.5	2.5	13.6	1.2	6.2	2.5	2.5
St. dev.	0.09	0.07	0.08	0.06	0.15	0.13	0.11	0.12	0.10
Root MSE	0.11	0.07	0.08	0.06	0.19	0.13	0.12	0.12	0.10
($\beta = 0.2$, $p^* = 0.8$)		$\gamma = 0.57$					$\alpha = 1.08$		
Mean	0.67	0.54	0.67	0.57	0.93	1.12	1.00	1.07	1.06
Mean bias (%)	17.5	5.3	17.5	0.0	13.9	3.7	7.4	0.9	1.8
St. dev.	0.17	0.10	0.17	0.09	0.21	0.13	0.12	0.16	0.11
Root MSE	0.20	0.11	0.20	0.09	0.26	0.13	0.15	0.16	0.11
($\beta = 0.5$, $p^* = 0.2$)		$\gamma = -2.2$					$\alpha = 2.6$		
Mean	−2.10	−2.24	−2.15	−2.19	2.15	2.70	2.38	2.59	2.58
Mean bias (%)	4.5	1.8	2.3	0.4	17.3	3.8	8.5	0.4	0.8
St. dev.	0.09	0.09	0.62	0.08	0.31	0.28	0.62	0.25	0.13
Root MSE	0.13	0.10	0.62	0.08	0.55	0.30	0.66	0.25	0.13
($\beta = 0.5$, $p^* = 0.5$)		$\gamma = -1.1$					$\alpha = 2.2$		
Mean	−0.98	−1.09	−1.08	−1.10	1.96	2.19	2.14	2.19	2.18
Mean bias (%)	10.9	0.9	1.81	0.0	10.9	0.4	2.7	0.4	0.9
St. dev.	0.11	0.09	0.14	0.06	0.21	0.17	0.11	0.16	0.10
Root MSE	0.16	0.09	0.14	0.06	0.32	0.17	0.13	0.16	0.10
($\beta = 0.5$, $p* = 0.8$)		$\gamma = -0.4$					$\alpha = 2.6$		
Mean	−0.06	−0.42	−0.21	−0.38	2.16	2.62	2.48	2.58	2.56
Mean bias (%)	85.0	5.0	47.5	5.0	16.9	0.8	4.6	0.8	1.5
St. dev.	0.25	0.10	0.34	0.10	0.31	0.13	0.15	0.24	0.12
Root MSE	0.42	0.10	0.39	0.10	0.54	0.13	0.19	0.24	0.13

100 replications. Root MSE denotes root mean squared error.

Namely, both GMM and ML perform well when the amount of state dependence is moderate, but GMM biases tend to be higher the higher the persistence in the data.

4. An application to female labour force participation

We illustrated the previous methods by estimating a relationship between female labour force participation and children variables allowing for individual effects. We

Table 4
Means and standard deviations of the estimators, $N = 1000, T = 6$

	GMM	ML	ML-MP	RML	GMM	ML	ML-MP	CML	RML
($\beta = 0.2,\ p^* = 0.2$)		$\gamma = -1.66$					$\alpha = 1.08$		
Mean	−1.64	−1.67	−1.70	−1.66	1.0	1.13	1.04	1.08	1.08
Mean bias (%)	1.2	0.6	2.4	0.0	7.4	4.6	3.7	0.0	0.0
St. dev.	0.05	0.04	0.13	0.04	0.15	0.11	0.09	0.12	0.07
Root MSE	0.06	0.05	0.13	0.05	0.17	0.12	0.10	0.12	0.07
($\beta = 0.2,\ p^* = 0.5$)		$\gamma = -0.4$					$\alpha = 0.81$		
Mean	−0.37	−0.40	−0.38	−0.39	0.76	0.81	0.77	0.80	0.79
Mean bias (%)	7.5	0.0	5.0	2.5	6.2	0.0	4.9	1.2	2.5
St. dev.	0.05	0.05	0.06	0.04	0.09	0.08	0.06	0.08	0.06
Root MSE	0.06	0.05	0.06	0.05	0.11	0.08	0.07	0.08	0.06
($\beta = 0.2,\ p^* = 0.8$)		$\gamma = 0.57$					$\alpha = 1.08$		
Mean	0.61	0.56	0.62	0.57	1.02	1.10	1.04	1.08	1.07
Mean bias (%)	7.0	1.7	8.8	0.0	5.5	1.8	3.7	0.0	0.9
St. dev.	0.12	0.08	0.09	0.07	0.16	0.10	0.10	0.13	0.09
Root MSE	0.13	0.08	0.11	0.07	0.17	0.10	0.11	0.13	0.09
($\beta = 0.5,\ p^* = 0.2$)		$\gamma = -2.2$					$\alpha = 2.6$		
Mean	−2.16	−2.22	−2.24	−2.20	2.40	2.65	2.55	2.60	2.60
Mean bias (%)	1.8	0.9	1.8	0.0	7.7	1.9	1.9	0.0	0.0
St. dev.	0.07	0.07	0.12	0.05	0.25	0.18	0.12	0.17	0.08
Root MSE	0.08	0.07	0.13	0.05	0.32	0.19	0.13	0.17	0.08
($\beta = 0.5,\ p^* = 0.5$)		$\gamma = -1.1$					$\alpha = 2.2$		
Mean	−1.02	−1.11	−1.07	−1.09	2.05	2.20	2.15	2.17	2.19
Mean bias (%)	7.3	0.9	2.7	0.9	6.8	0.0	2.3	1.4	0.4
St. dev.	0.08	0.08	0.12	0.04	0.15	0.14	0.08	0.11	0.06
Root MSE	0.11	0.08	0.12	0.05	0.21	0.14	0.10	0.11	0.06
($\beta = 0.5,\ p^* = 0.8$)		$\gamma = -0.4$					$\alpha = 2.6$		
Mean	−0.24	−0.41	−0.31	−0.39	2.39	2.61	2.55	2.60	2.60
Mean bias (%)	40.0	2.5	22.5	2.5	8.1	0.4	1.9	0.0	0.0
St. dev.	0.19	0.08	0.14	0.06	0.25	0.11	0.1	0.18	0.09
Root MSE	0.25	0.08	0.16	0.06	0.33	0.11	0.12	0.18	0.09

100 replications. Root MSE denotes root mean squared error.

used data on 384 white married women from the random sub-sample of the Panel Study of Income Dynamics (PSID), for the years 1971, 1973, 1975, and 1977. Only women continuously married with the same husband and who were 20–50 years old in 1971 were included in the sample.

The starting point is an equation of the form:

$$y_{it} = \mathbf{1}(\gamma + \beta' x_{it} + \eta_i + v_{it} \geqslant 0), \tag{4.1}$$

where $y_{it}=1$ if the ith woman worked in year t. The effect of children is specified by the vector x_{it} which consists of two dummy variables: $x_{it} = (k12_{it}, k35_{it})'$. The first dummy equals 1 if the age of the youngest child is 1 or 2, while the second takes the value 1 if the youngest child is aged 3, 4 or 5. This particular specification is motivated by the fact

Table 5
Means and standard deviations of fixed effects estimates of α $N = 500$, $\beta = 0.2$

	$T = 4$		$T = 6$
$p^* = 0.2$		$\alpha = 1.08$	
Mean	−1.42		−0.30
St. dev.	0.21		0.13
Root MSE	2.51		1.39
$p^* = 0.5$		$\alpha = 0.81$	
Mean	−1.32		−0.26
St. dev.	0.18		0.11
Root MSE	2.14		1.08
$p^* = 0.8$		$\alpha = 1.08$	
Mean	−1.40		−0.27
St. dev.	0.23		0.13
Root MSE	2.49		1.36

100 replications. Root MSE denotes root mean squared error.

that most of the children's effects on participation appear to depend on the presence of very young children, more so than, for example, on the total number of children living in the household (see Browning, 1992). The individual-specific effect η_i will capture unobserved permanent components in both wages and tastes for non-market time. [13]

Table 6 contains the estimates of four different logit specifications of the basic model that treat children as strictly exogenous variables, with and without time dummies. For simplicity we describe the methods with reference to the latter. Column a labelled "pooled levels" presents the results from a model without unobserved heterogeneity. In this case the log-likelihood function takes the form

$$L_a = \sum_{i=1}^{N}\sum_{t=1}^{T}\{y_{it}\ln F_{it} + (1 - y_{it})\ln(1 - F_{it})\}, \tag{4.2}$$

where $F_{it} = F(\gamma + \beta' x_{it})$.

Column b reports estimates from a pseudo-conditional logit log-likelihood that leaves the distribution of the effects unrestricted. The form of the criterion is

$$L_b = \sum_{t=2}^{T}\sum_{i=1}^{N}\{y_{it}(1 - y_{i(t-1)})\ln F(\beta'\Delta x_{it}) + (1 - y_{it})y_{i(t-1)}\ln[1 - F(\beta'\Delta x_{it})]\} = \sum_{t=2}^{T} L_{bt}. \tag{4.3}$$

[13] Hyslop (1999) estimated female participation equations with state dependence, serial correlation and individual effects, using a richer set of explanatory variables but ruling out feedback from participation histories into fertility decisions.

Table 6
Female labour force participation logit models, exogenous children

Independent variables	Pseudo-ML			GMM
	a	b	c	d
	Pooled levels	Conditional logit	Linear effects	Unrestricted effects
Models without time dummies				
$k12_t$	−1.74	−1.62	−1.26	−1.62
	(0.26)	(0.71)	(0.31)	(0.49)
$k35_t$	−0.60	−1.27	−0.76	−0.92
	(0.17)	(0.50)	(0.18)	(0.08)
Constant	0.42		0.38	0.44
	(0.06)		(0.10)	(0.07)
Models with time dummies				
$k12_t$	−1.70	−1.38	−1.14	−1.37
	(0.27)	(0.66)	(0.29)	(0.51)
$k35_t$	−0.56	−1.06	−0.67	−0.71
	(0.17)	(0.53)	(0.18)	(0.12)
$Constant_{71}$	0.41		0.36	0.38
	(0.11)		(0.12)	(0.06)
$Constant_{73}$	0.27	0.53	0.23	0.29
	(0.11)	(0.27)	(0.12)	(0.06)
$Constant_{75}$	0.40	0.52	0.38	0.43
	(0.11)	(0.27)	(0.13)	(0.11)
$Constant_{77}$	0.56	−0.28	0.56	0.56
	(0.11)	(0.25)	(0.13)	(0.05)

$N = 384$, white married women between 20 and 50 years old in 1971, from the PSID random subsample. Years = 1971, 1973, 1975, 1977.
k12 = 1 if the age of the youngest child is 1 or 2.
k35 = 1 if the age of the youngest child is 3, 4 or 5.
Figures in parentheses are standard errors.
Estimated constants for conditional logit with time dummies are in first differences.

The term L_{bt} is the conditional logit log-likelihood given a sufficient statistic for the fixed effect of a panel consisting of waves $t-1$ and t only (Chamberlain, 1980). Thus, although the estimator that maximizes L_b is consistent regardless of the value of T, L_b would only be the actual conditional logit log-likelihood when $T = 2$.

Column c reports pseudo-ML estimates from a model in which the conditional mean of the effects is restricted to be linear (as in Chamberlain, 1984). The form of the criterion in this case is the same as (4.2), but with probabilities given by

$$F_{it} = F\left(\gamma + \beta' x_{it} + \sum_{s=1}^{T} \lambda_s' x_{is}\right). \tag{4.4}$$

Finally, column d contains GMM estimates of a generalization of the previous model that leaves the conditional mean of the effects unrestricted (cf. Newey, 1994a).

Estimation is based on the following sample orthogonality conditions:

$$b_{tN} = \sum_i \mathbf{1}(x_i^T = \phi_j^T)(g_{it}^* - \beta' x_{it}^*), \tag{4.5}$$

where $g_{it} = F^{-1}[\hat{Pr}(y_t = 1 \mid x_i^T)], \hat{Pr}(y_t = 1 \mid x_i^T)$ are cell-specific sample frequencies, and the stars denote that the variables have been transformed into orthogonal deviations (cf. Arellano and Bover, 1995, and further detail given below). All the cells with less than four observations were dropped, and as a result the number of orthogonality conditions used in the estimation was also reduced.

We can observe that relative to the rest of the estimates, the pooled levels estimates of the coefficient on k12 are larger (in absolute value) while those of the coefficient on k35 are smaller. Aside from this, it is of some interest to compare the conditional effects estimates in column b, with the random effects estimates in columns c and d. The estimates in column c, which are the most restrictive, are more at variance with the conditional effects estimates than the less restrictive estimates shown in column d. Note, however, that models c and d are not nested within model b, because model b assumes that $v_{it} \mid x_i^T, \eta_i$ is logistic whereas models c and d assume that it is the composite error $\eta_i + v_{it} \mid x_i^T$ the one which has a logistic distribution.

The previous remarks are true for both the estimates with and without time dummies, although the latter are smaller than the former in all cases. This fact suggests the presence of non-negligible cyclical effects on female participation.

Table 7 contains GMM estimates that treat the children variables as predetermined by conditioning on lagged children and lagged participation. The estimates are based on orthogonality conditions of the type described in Section 2, except for the fact that we used orthogonal deviations as opposed to first-differences. Specifically, we used the following sample moments

$$b_{tN}^d = \sum_i z_{it}(f_{it}^* - \gamma_t^* - \beta' x_{it}^*) \quad (t = 2,3),$$

$$b_{tN}^\ell = \sum_i (f_{it} - \gamma_t - \beta' x_{it}) \quad (t = 2,3,4),$$

where $f_{it} = F^{-1}[\widehat{\Pr}(y_t = 1 \mid y_i^{t-1}, x_i^t)]$, and starred variables denote orthogonal deviations as in $f_{i2}^* = \sqrt{2/3}[f_{i2} - 0.5(f_{i3} + f_{i4})]$ and $f_{i3}^* = \sqrt{1/2}(f_{i3} - f_{i4})$. Of a total of 32 possible values of (y_{i1}, x_{i1}, x_{i2}) and 256 of $(y_{i1}, y_{i2}, x_{i1}, x_{i2}, x_{i3})$, only 17 and 37 occurred in the data. Hence z_{i2} and z_{i3} are, respectively, 17×1 and 37×1 vectors containing binary indicators for those outcomes. Moreover, we only calculated unrestricted frequencies $\widehat{\Pr}(y_t = 1 \mid y_i^{t-1}, x_i^t)$ for cells containing at least 4 observations. In this way, the numbers of cell frequencies used by the GMM estimator were 9 for (y_{i1}, x_{i1}, x_{i2}), 12 for $(y_{i1}, y_{i2}, x_{i1}, x_{i2}, x_{i3})$ and 13 for $(y_{i1}, y_{i2}, y_{i3}, x_{i1}, \ldots, x_{i4})$, and the effective sample size was reduced from 384 to 308.[14] With these specifications, we calculated a one-step

[14] Except for b_{2N}^ℓ and b_{3N}^ℓ which were based on 360 and 339 observations, respectively.

Table 7
Female labour force participation logit models, predetermined children

Independent variables	GMM			
	a	b	c	d
$k12_t$	−3.14	−3.26	−2.00	−2.18
	(0.64)	(0.65)	(0.60)	(0.61)
$k35_t$	−1.40	−1.15	−0.77	−0.61
	(0.35)	(0.40)	(0.36)	(0.39)
y_{t-1}		1.41		1.20
		(0.53)		(0.55)
Constant	0.64	−0.15		
	(0.14)	(0.34)		
$Constant_{73}$			0.37	−0.29
			(0.17)	(0.34)
$Constant_{75}$			0.61	−0.04
			(0.18)	(0.34)
$Constant_{77}$			0.81	0.09
			(0.20)	(0.42)

$N = 384$, white married women between 20 and 50 years old in 1971, from the PSID random subsample. Years = 1971, 1973, 1975, 1977.

k12 = 1 if the age of the youngest child is 1 or 2.

k35 = 1 if the age of the youngest child is 3, 4 or 5.

Figures in parentheses are standard errors.

GMM estimator of the 5 parameters $(\gamma_2, \gamma_3, \gamma_4, \beta')$, based on 57 moments, 34 sample frequencies, and 308 observations, that maximized

$$\sum_{t=2}^{3} b_{tN}^{d\prime} \left(\sum_i z_{it} z_{it}' \right)^{-1} b_{tN}^{d} + \sum_{t=2}^{4} b_{tN}^{\ell\prime} b_{tN}^{\ell}.$$

We report estimates with and without period-specific intercepts, for both the basic model and an extended model that includes lagged participation as a regressor. Interestingly, the baseline model's estimates are markedly different from the estimates obtained in Table 6 under strict exogeneity, implying stronger effects of small children on participation. The interpretation of these differences, however, is not straightforward since they may be the result of misspecification.[15] In fact, allowing for variation over time in the intercepts substantially reduces the impact of the children variables. Finally, when time dummies are included, lagged participation is found to be marginally significant. We also tried a more general specification including interactions between

[15] Nevertheless, it may be mentioned that, working with a different framework, Rosenzweig and Wolpin (1980) found that the use of actual fertility in participation equations understated the impact of exogenous changes in fertility on female work status.

Table 8
Effects on female labour force participation probabilities logit models, predetermined children

	$\beta = (-3.14, -1.40)$			$\beta = (-2.00, -0.77)$		
	1973	1975	1977	1973	1975	1977
Changing (k12, k35)						
from (0,1) to (1,1)	−0.33	−0.36	−0.38	−0.29	−0.30	−0.30
from (0,0) to (1,0)	−0.46	−0.46	−0.45	−0.31	−0.30	−0.29
from (1,0) to (1,1)	−0.09	−0.11	−0.12	−0.10	−0.11	−0.11

lagged participation and children, but none of the interactions were significant, and the other coefficients did not change.

In order to obtain consistent standard errors for the GMM estimates in Table 7 and in column d of Table 6, we estimated the variance of the sample orthogonality conditions, and took into account that the weighting matrix was non-optimal. This involved the estimation of the joint covariance matrix of the moment restrictions using the true probabilities, and the unrestricted cell sample frequencies (see Appendix A).

The implications of the results in Table 7 are that young children have a negative effect on female labour participation (from the signs of the coefficients) and that 1–2 year olds have a larger negative effect than 3–5 year olds (from the ratio between the estimates). To acquire additional information of interest we calculated the implied structural changes in the probabilities using the marginal effect estimator described in (2.55). This estimator holds constant the indirect effect of children on participation due to their dependence with individual effects. Table 8 reports changes in the participation probabilities corresponding to estimates with and without year effects in columns a and c of Table 7. Controlling for year effects, having a 1–2 year old child reduces the probability of participation by approximately 30 percentage points, while having a 3–5 year old reduces it by 10 percentage points.

The reported estimates are calculated under the assumption that variances are constant over time. The impact of allowing for unequal variances on the estimated children effects is an issue that remains to be explored.

Simulation evidence on the properties of the estimators. We simulated data calibrated to the PSID sample to study the finite sample properties of the GMM estimator in the empirical application. This also had the additional interest of exhibiting some Monte Carlo results with heterogeneity, which complement those reported in Section 3 and illustrate how to generate heterogeneous data from our model.

In order to specify the data generation process, we first have to choose the values of the structural parameters β in (4.1) and of $\psi_j^T = \mathrm{E}(\eta \mid w^T = \phi_j^T)$. In our case $T = 4$ and $w^T = (x_1, \ldots, x_4, y_1, \ldots, y_3)$ is an 11×1 vector of binary variables, so that there are $2^{11} = 2048$ different ψ_j^4. We also need to choose $\Pr(x_1 = \zeta_j^2, y_1 = \zeta_k^1)$ and $\pi_{tj}(x^{t-1}, y^{t-1}) = \Pr(x_t = \zeta_j^2 \mid x^{t-1}, y^{t-1})$ for $t = 2, 3, 4$ and each possible value of (x^{t-1}, y^{t-1}). Next, given these quantities, we obtain $\mathrm{E}(\eta \mid w^3)$ for given $w^3 = (x_1, x_2, x_3, y_1, y_2)$ as the solution to

the non-linear equation:

$$\begin{aligned}\mathrm{E}(\eta \mid w^3) &= \sum_j \sum_k \mathrm{E}(\eta \mid w^3, x_4 = \zeta_j^2, y_3 = \zeta_k^1) \Pr(x_4 = \zeta_j^2, y_3 = \zeta_k^1 \mid w^3) \\ &= \sum_j \sum_k \mathrm{E}(\eta \mid w^3, x_4 = \zeta_j^2, y_3 = \zeta_k^1) \\ &\quad \times \Pr(x_4 = \zeta_j^2 \mid w^3, y_3 = \zeta_k^1) \Pr(y_3 = \zeta_k^1 \mid w^3),\end{aligned}$$

where $\Pr(y_3 = \zeta_k^1 \mid w^3)$ is the probability specified by the model, e.g.

$$\Pr(y_3 = 1 \mid w^3) = F[\beta' x_3 + \mathrm{E}(\eta \mid w^3)].$$

The terms $\mathrm{E}(\eta \mid w^3, x_4 = \zeta_j^2, y_3 = \zeta_k^1)$ and $\Pr(x_4 = \zeta_j^2 \mid w^3, y_3 = \zeta_k^1)$ correspond to those labelled above as ψ_j^4 and π_{4j}, respectively. The procedure is repeated $2^8 = 256$ times to obtain as many $\psi_j^3 = \mathrm{E}(\eta \mid w^3 = \phi_j^3)$. Finally, we obtain the $2^5 = 32$ terms $\psi_j^2 = \mathrm{E}(\eta \mid w^2 = \phi_j^2)$ solving similar non-linear equations as functions of ψ_j^3 and π_{3j}.

Having specified the model, we can begin by simulating data on (x_1, y_1) with probabilities $\Pr(x_1 = \zeta_j^2, y_1 = \zeta_k^1)$. Next, we obtain data on x_2 from $\pi_{2j}(x_1, y_1)$, and then on y_2 using the model's probability $F[\beta' x_2 + \mathrm{E}(\eta \mid w^2)]$. These are followed by data on x_3 from $\pi_{3j}(x_2, x_1, y_1, y_2)$, and so on.

We set β to the estimated values in Table 7, col. 1, $\beta = (-3.14, -1.40)$. To select values for ψ_j^4 we considered a linear specification of the conditional mean by estimating logit equations of the form

$$y_{i4} = 1(-3.14x_{1i4} - 1.4x_{2i4} + \gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{i2} + \gamma_3 y_{i3} + \gamma_4' x_i^T + \varepsilon_{i4} > 0),$$

and set $\mathrm{E}(\eta \mid w^4) = -1.3 + 0.49 y_{i1} + 0.89 y_{i2} + 2.6 y_{i3}$, which correspond to the empirical estimates of γ_0, γ_1 , γ_2 and γ_3 with $\gamma_4 = 0$, since the estimated γ_4 were insignificant. Initial observations of x_1 and y_1 were randomly generated using their marginal probabilities in the empirical data (0.10, 0.16, and 0.54, respectively). Subsequent observations of x_2, x_3 and x_4 were generated from univariate unrestricted autoregressive logit equations since cross terms were mostly insignificant. [16]

[16] The equations are

$$x_{1i2} = 1(-3.0 + 2.3x_{1i1} + \xi_{1i2} > 0),$$

$$x_{1i3} = 1(-4.2 - 0.03x_{1i1} + 3.8x_{1i2} + \xi_{1i3} > 0),$$

$$x_{1i4} = 1(-4.7 + 0.88x_{1i1} + 0.36x_{1i2} + 3.3x_{1i3} + \xi_{1i4} > 0),$$

$$x_{2i2} = 1(-2.5 + 2.0x_{2i1} + \xi_{2i2} > 0),$$

$$x_{2i3} = 1(-2.9 - 0.97x_{2i1} + 2.3x_{2i2} + \xi_{2i3} > 0),$$

$$x_{2i4} = 1(-3.8 + 1.8x_{2i1} - 0.58x_{2i2} + 2.0x_{2i3} + \xi_{2i4} > 0).$$

Table 9
Monte Carlo simulation for the GMM estimates used in the empirical application $N = 384$, $T = 4$

	$\beta_1 = -3.14$	$\beta_2 = -1.41$
Mean	−3.10	−1.41
St. dev.	1.43	0.43
Skewness	3.9	0.35
Kurtosis	24.5	3.7
Quantiles:		
0.10	−4.18	−1.93
0.25	−3.76	−1.69
0.50	−3.28	−1.45
0.75	−2.68	−1.14
0.90	−2.15	−0.88

100 replications.

Table 9 contains the simulation results for the GMM estimates used in the first column of Table 7. The results for this experiment are encouraging, except for the fact that with $N = 384$ there is some evidence of non-normality in the sampling distribution of the estimate of β_1.[17] Both mean and median biases are negligible, and measured dispersion indicates that the estimates are reasonably informative for the sample size used in the application. These results, however, should be viewed with caution since with such a small sample size the variances of the estimates are likely to be sensitive to alternative specifications of the conditional mean of the effects and the processes of the explanatory variables.

Acknowledgements

We are grateful to Miguel Delgado, Jim Heckman, Pedro Mira, Franco Peracchi, Alain Trognon, Enrique Sentana, an associate editor, two anonymous referees, and participants at the First Bergamo Workshop on Applied Economics, 11–12 October 1996, for helpful comments on this work. All remaining errors are our own. The first author acknowledges research funding from the Spanish DGES, Grant PB96-0134.

Appendix A. Formulae for standard errors

All the GMM estimates that we use in the empirical application in Section 4 can be written in the general form

$$\hat{\theta} = \arg\min{}' \left[\frac{1}{N} \sum_{i=1}^{N} \psi_i(\hat{p}, \theta) \right]' A_N \left[\frac{1}{N} \sum_{i=1}^{N} \psi_i(\hat{p}, \theta) \right],$$

[17] In a simulation of the same model with $N = 3000$ (not reported) the skewness and kurtosis coefficients of the estimate of β_1 were −0.2 and 3.2, respectively.

where A_N is a weight matrix, and $\hat{p}$ is a vector of cell-specific sample frequencies that consistently estimate the unknown probabilities p. We may regard $\hat{p}$ as the solution to the moment equations

$$\frac{1}{N}\sum_{i=1}^{N} h_i(\hat{p}) = 0,$$

where $h_i(.)$ is a vector of linear functions of the same dimension as p, so that

$$\sqrt{N}(\hat{p} - p) = \left(-\frac{1}{N}\sum_{i=1}^{N} \frac{\partial h_i(p)}{\partial p'} \right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} h_i(p).$$

Moreover, using a first-order Taylor expansion we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi_i(\hat{p},\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi_i(p,\theta) + \left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \psi_i(p,\theta)}{\partial p'} \right) \sqrt{N}(\hat{p} - p) + o_p(1)$$

$$= [I, -Q_N] \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i(p,\theta) + o_p(1),$$

where $\zeta_i(p,\theta) = [\psi_i(p,\theta)', h_i(p)']'$ and $Q_N = (\sum_i \partial\psi_i(p,\theta)/\partial p')(\sum_i \partial h_i(p)/\partial p')^{-1}$. Thus a consistent estimate of the asymptotic variance of $N^{-1/2}\sum_i \psi_i(\hat{p},\theta)$ is

$$\hat{W} = [I, -\hat{Q}_N] \frac{1}{N}\sum_{i=1}^{N} \zeta_i(\hat{p},\hat{\theta})\zeta_i(\hat{p},\hat{\theta})'[I, -\hat{Q}_N]',$$

where $\hat{Q}_N$ is similar to Q_N but replacing p and θ by $\hat{p}$ and $\hat{\theta}$, respectively. Finally, from standard GMM theory, a consistent estimator of the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta)$ is given by the sandwich formula

$$\hat{V}_\theta = (\hat{D}_\theta' A_N \hat{D}_\theta)^{-1} \hat{D}_\theta' A_N \hat{W} A_N \hat{D}_\theta (\hat{D}_\theta' A_N \hat{D}_\theta)^{-1},$$

where $\hat{D}_\theta = N^{-1}\sum_{i=1}^{N} \partial\psi_i(\hat{p},\hat{\theta})/\partial\theta'$.

References

Amemiya, T., 1985. Advanced Econometrics. Basil Blackwell, Oxford.

Andersen, E., 1970. Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, Series B 32, 283–301.

Anderson, T.W., Hsiao, C., 1981. Estimation of dynamic models with error components. Journal of the American Statistical Association 76, 598–606.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58, 277–297.

Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. Journal of Econometrics 68, 29–51.

Arellano, M., Honoré, B., 2001. Panel data models: some recent developments. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, Vol. 5. North-Holland, Amsterdam.

Beasley, J.D., Springer, S.G., 1977. The percentage points of the normal distribution. Applied Statistics 26, 118–121.

Berkson, J., 1944. Application of the logistic function to bio-assay. Journal of the American Statistical Association 39, 357–365.
Bover, O., Arellano, M., Bentolila, S., 2002. Unemployment duration, benefit duration, and the business cycle. Economic Journal 112, 223–265.
Browning, M., 1992. Children and household economic behavior. Journal of Economic Literature 30, 1434–1475.
Card, D., Sullivan, D., 1988. Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 56, 497–530.
Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economic Studies 47, 225–238.
Chamberlain, G., 1984. Panel data. In: Griliches, Z., Intrilligator, M.D., (Eds.), Handbook of Econometrics, Vol. 2. North-Holland, Amsterdam.
Chamberlain, G., 1985. Heterogeneity, omitted variable bias, and duration dependence. In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge, UK.
Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34, 305–334.
Chen, S., 1998. Root-*N* consistent estimation of a panel data sample selection model. Unpublished manuscript, The Hong Kong University of Science and Technology.
Cox, D.R., 1970. Analysis of Binary Data. Methuen, London.
Delgado, M.A., Mora, J., 1995. Nonparametric and semiparametric estimation with discrete regressors. Econometrica 63, 1477–1484.
Heckman, J.J., 1981a. Statistical models for discrete panel data. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.
Heckman, J.J., 1981b. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.
Heckman, J.J., 1981c. Heterogeneity and state dependence. In: Rosen, S. (Ed.), Studies in Labor Markets. NBER, University of Chicago Press, Chicago.
Heckman, J.J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.
Holtz-Eakin, D., Newey, W., Rosen, H.S., 1988. Estimating vector autoregressions with panel data. Econometrica 56, 1371–1395.
Honoré, B., Kyriazidou, E., 2000. Panel data discrete choice models with lagged dependent variables. Econometrica 68, 839–874.
Honoré, B., Lewbel, A., 2002. Semiparametric binary choice panel data models without strictly exogeneous regressors. Econometrica 70, 2053–2063.
Hyslop, D.R., 1999. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. Econometrica 67, 1255–1294.
Magnac, T., 2000. State dependence and unobserved heterogeneity in youth employment histories. Economic Journal 110, 805–837.
Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55, 357–362.
Moon, C.-G., Stotsky, J.G., 1993. The effect of rent control on housing quality change: a longitudinal analysis. Journal of Political Economy 101, 1114–1148.
Newey, W., 1994a. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.
Newey, W., 1994b. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10, 233–253.
Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. 4. North Holland, Amsterdam.
Rosenzweig, M.R., Wolpin, K.I., 1980. Life-cycle labor supply and fertility: causal inferences from household models. Journal of Political Economy 88, 328–348.