CAM

**Centre for Applied
Microeconometrics**

**Institute of Economics
University of Copenhagen**

http://www.econ.ku.dk/CAM/

Bounds on Parameters in Dynamic Discrete Choice Models

Bo E. Honoré and Elie Tamer

2004-23

**The activities of CAM are financed by a grant from
The Danish National Research Foundation**

Bounds on Parameters in Dynamic Discrete Choice Models[*]

Bo E. Honoré
Department of Economics
Princeton University

Elie Tamer
Department of Economics
Princeton University

First version: October 2002. This version: August 2004.

Abstract

Identification of dynamic nonlinear panel data models is an important and delicate problem in econometrics. In this paper we provide insights that shed light on the identification of parameters of some commonly used models. Using this insight, we are able to show through simple calculations that point identification often fails in these models. On the other hand, these calculations also suggest that the model restricts the parameter to lie in a region that is very small in many cases, and the failure of point identification may therefore be of little practical importance in those cases. Although the emphasis is on identification, our techniques are constructive in that they can easily form the basis for consistent estimates of the identified sets.

1 Introduction

Dynamic panel data models have played an important role in applied economics dating back to the work of Balestra and Nerlove (1966). Econometric specifications of these models typically specify features of the conditional distribution of the dependent variable of interest for an individual i, y_{it}, conditional on lagged values of that variable, a set of possibly time-varying explanatory variables, x_i, and on an individual specific unobserved variable, α_i. To fully specify such a model, one needs to

[*] Mailing addresses: Princeton University, Department of Economics, Princeton NJ 08544-1021. The authors gratefully acknowledge financial support from the National Science Foundation, The Gregory C. Chow Econometric Research Program at Princeton University, and the Danish National Research Foundation (through CAM at The University of Copenhagen). We thank Luojia Hu, Nandita Gawade and seminar participants at the University of Copenhagen, UCL, UCLA, UC Santa Barbara, UC Berkeley, Maryland, NYU, Northwestern, Rice, Columbia, Cornell and the NBER-NSF Conference on Panel Data Analysis for helpful comments.

also specify the distribution of the individuals' dependent variable in the initial period conditional on those variables. This is delicate if the process for an individual started prior to the initial period in the sample because the distribution of the first observation will be tied to the distribution of the later observations in a way that depends on what one assumes about how the process was started and on the evolution of the explanatory variables prior to the sampling period.

For example, a parametric model might specify that the conditional distribution of y_{it} depends on lagged values only though y_{it-1}, in which case the conditional distribution of y_{it} has the form

$$f_t\left(y_{it}|\, y_{it-1}, x_i, \alpha_i; \theta\right) \tag{1}$$

where θ is a vector of unknown parameters to be estimated. The vector, x_i, can consist of variables that are constant over time as well as of the entire sequence of time–varying explanatory variables. In a fully parametric (random effects) approach, one specifies the distribution of α_i conditional on the explanatory variables x_i. In practice, α_i is frequently assumed to be independent of x_i, and the random effects approach then specifies the distribution of α_i. If (1) is static in the sense that the density does not depend on y_{it-1}, then this would allow one to write the likelihood function for an individual with T observations as

$$\begin{aligned} \mathcal{L} &= \int \prod_{t=1}^{T} f_t\left(y_{it}|\, y_{it-1}, x_i, \alpha_i\right) f\left(\alpha_i|\, x_i\right) d\alpha_i \\ &= \int \prod_{t=1}^{T} f_t\left(y_{it}|\, x_i, \alpha_i\right) f\left(\alpha_i|\, x_i\right) d\alpha_i \end{aligned}$$

On the other hand, if the model is dynamic so (1) does depend on y_{it-1}, the likelihood function has the structure

$$\mathcal{L} = \int f_1\left(y_{i1}|\, x_i, \alpha_i\right) \prod_{t=2}^{T} f_t\left(y_{it}|\, y_{it-1}, x_i, \alpha_i\right) f\left(\alpha_i|\, x_i\right) d\alpha_i$$

Unfortunately, it is not clear how one would go from (1) to $f_1\left(y_{i1}|\, x_i, \alpha_i\right)$ since the relationship between the two typically depends on the evolution of the explanatory variables before the start of the sampling period. This is what is known as the *initial conditions problem.* Alternatively, one could work with the likelihood function conditional on the first observation, y_{i1}. This often leads to convenient functional forms. But the random effects approach can be problematic in this case if one wants it to be internally consistent across different number of time periods[1].

[1] For a discussion of these issues see Wooldridge (2000) and Honoré (2002).

A so–called fixed effects approach, on the other hand, attempts to estimate θ without making any assumptions on $f(\alpha_i | x_i)$. This will in principle circumvent the initial conditions problem, but there are at least two other problems with this approach. First, there are many dynamic panel data models for which it is not known how to do fixed effects estimation, and even when it is known, the maintained assumptions are often very strong. For example, the estimator of the dynamic discrete choice model proposed by Honoré and Kyriazidou (2000) requires one to "match" the explanatory variables in different time–periods, which rules out time–specific effects. Secondly, as discussed by Wooldridge (2000), knowing θ does not always allow one to calculate the "marginal" effect of interest.

The point of departure for this paper is that the random effects approach is attractive because it allows one to estimate all quantities of interest, but that specifying $f(\alpha_i | x_i, y_{i1})$ or $f_1(y_{i1} | x_i, \alpha_i)$ can be problematic. The contribution of this paper is to provide insights through simple calculations that allow us to examine the identified features of these models without making any assumptions on $f(\alpha_i | x_i, y_{i1})$ or $f_1(y_{i1} | x_i, \alpha_i)$. Our calculations show that the parameters of interest are not point identified in simple commonly used models. However the size of the identified regions suggest that this lack of identification may not be of great practical importance. Although the emphasis is on identification, our techniques are constructive in that they can easily be used to obtain consistent estimates of the identified sets.

To focus ideas, we concentrate on special cases of the dynamic random effects discrete choice model

$$y_{it} = 1\left\{x_{it}'\beta + y_{i,t-1}\gamma + \alpha_i + \varepsilon_{it} \geq 0\right\} \tag{2}$$

Recent empirical applications of this model include studies of labor force participation (Hyslop (1999)), health status (Contoyannis, Jones, and Rice (2002) and Halliday (2002)), product purchase behavior (Chintagunta, Kyriazidou, and Perktold (2001)) and welfare participation (Chay, Hoynes, and Hyslop (2001)). Honoré and Kyriazidou (2000) showed how to estimate and identify a fixed effects version of this model in the case where the change in the vector of explanatory variables has support in a neighborhood around 0. One of the contributions of this paper is to show that such an assumption is often necessary for identification, but that it is sometimes possible to construct tight bounds on the parameters of such a model, even when it is not known that the model is identified. A number of other papers have also proposed estimation of the parameters of versions of (2). For example, Arellano and Carrasco (2003) explicitly model the distribution of $\alpha_i + \varepsilon_{it}$ conditional on current and past observed values of x_{it} and on the observed past values of y_{it}. Our aim here is to study what can be learned about parameters of interest without such assumptions.

Throughout this paper we will consider situations where the data consists of $\{(y_{it}, x_{it})\}_{t=1}^{T}$ and we use the notation $w_i^t = (w_{i1}, ..., w_{it})$. K will denote the number of elements in x_i^T.

2 Basic Ideas

2.1 Identification

Consider (1) augmented with a model for the individual specific effect α (given x^T), and let θ be the parameters of the model. In the dynamic probit model with normally distributed individual specific effects, θ would be $\left(\beta', \gamma, \sigma^2\right)$, where σ^2 is the variance of the individual specific effect. This model is incomplete in the sense that it does not allow one to calculate the distribution of the dependent variables conditional on the observed explanatory variables. For that, one would also need the distribution of the initial conditions, $p_0\left(\alpha, x^T\right) = P\left(y_0 = 1 | \, x^T, \alpha\right)$. More generically, we will use $p_0\left(\alpha, x^T\right)$ to denote the distribution of y_{i0} given $\left(x^T, \alpha\right)$.

One can think of $(p_0(\cdot,\cdot), \theta)$ as the unknown parameters of the model, and knowledge of $(p_0(\cdot,\cdot), \theta)$ allows one to calculate the conditional probability of any sequence of events A. In a discrete choice model, A will be any sequence of subsets of $\{0,1\}$ such that $A = A_1 \cap A_2 \cap \ldots A_T$. Let $\pi\left(A | \, x^T, \alpha; p_0(\cdot,\cdot), \theta\right)$ denote the probability of the event A given $\left(x^T, \alpha\right)$ predicted by the model:

$$\pi\left(A | \, x^T; p_0(\cdot,\cdot), \theta\right) = \int \pi\left(A | \, x^T, \alpha; p_0(\cdot,\cdot), \theta\right) dG\left(\alpha | \, x^T; \theta\right)$$

and let $G\left(\alpha | \, x^T; \theta\right)$ denote the distribution of α given x^T when the parameter value is θ. We use $\pi\left(\mathcal{A} | \, x^T; p_0(\cdot,\cdot), \theta\right)$ and $P\left(\mathcal{A} | \, x^T\right)$ to denote the set of all $\pi\left(A; p_0(\cdot,\cdot), \theta | \, x^T\right)$ and $P\left(A | \, x^T\right)$ where $A \in \mathcal{A}$. Here, $P\left(A | \, x^T\right)$ denotes the true conditional probability of A given x^T. In the dynamic discrete choice model $\pi\left(\mathcal{A}; p_0\left(\cdot, x^T\right), \theta | \, x^T\right)$ and $P\left(\mathcal{A} | \, x^T\right)$ are 2^T–dimensional vectors.

With this notation, the set of $(p_0(\cdot,\cdot), \theta)$ that are consistent with a particular data–generating process with probabilities $P\left(\mathcal{A} | \, x^T\right)$, is given by

$$\Psi = \left\{(p_0(\cdot,\cdot), \theta) : P\left(\pi\left(\mathcal{A} | \, x^T; p_0(\cdot,\cdot), \theta\right) = P\left(\mathcal{A} | \, x^T\right)\right) = 1\right\}$$

and the sharp bound on θ is given by

$$\Theta = \left\{\theta : \exists p_0(\cdot,\cdot) : \Re^{1+K} \to [0,1] \text{ such that } P\left(\pi\left(\mathcal{A} | \, x^T; p_0(\cdot,\cdot), \theta\right) = P\left(\mathcal{A} | \, x^T\right)\right) = 1\right\}$$

2.2 Calculation of the identified region

There are a number of ways to write the identified region as the solution to a minimization problem. We suggest three methods that can be used to obtain Θ: a minimum distance method, a maximum likelihood method, and a linear programming method. The latter is especially convenient and practical in the case where have discrete covariates. Applying the analogy principle to each of these will lead to a different estimator of the identified region.

2.2.1 Minimum Distance

One can write Ψ as the solution to the following "minimum distance" minimization problem

$$\min_{p_0(\cdot,\cdot),\theta} E\left[\left\|\pi\left(\mathcal{A}|\, x^T; p_0(\cdot,\cdot),\theta\right) - P\left(\mathcal{A}|\, x^T\right)\right\| w\left(x^T\right)\right] \qquad (3)$$

for some positive weighting function w. Conceptually, it is possible to obtain the set of parameters that solve the above minimization problem. In section 4, we provide some examples to illustrate this.

It is also clear that using a sample analog, one can obtain consistent estimates of the identified set. To get these estimates, one can use the analogy principle to obtain the empirical analog of (3). This entails obtaining consistent estimates of $P\left(\mathcal{A}|\, x_i^T\right)$ in a first step, and then collecting all the parameter values that come within ϵ_n from minimizing the sample objective function, where $\epsilon_n > 0$, and $\epsilon_n \to 0$ as sample size increases. For more on this see Manski and Tamer (2002). When the distribution of x is continuous, $P\left(\mathcal{A}|\, x_i^T\right)$ will be imprecisely estimated in regions where the density of x is small. The weighting function in (3) can be used to downweight the observations in this region.

2.2.2 Maximum likelihood

The identified set can also be characterized as maximizing a likelihood function. Recall that y_i denotes $\{y_{i1}, ..., y_{iT}\}$, and let d denote a sequence of T zeros or ones. Consider 2^T non–negative functions $g_d(\cdot)$ such that $\sum_d g_d(\cdot) = 1$. Standard theory for maximum likelihood implies that for any positive weighting function, w, the function

$$E\left[\log\left(g_{y_i}\left(x_i^T\right)\right) w\left(x_i^T\right)\right]$$

is uniquely maximized (as a function of the $g_d(\cdot)$'s) at

$$g_d\left(x_i^T\right) = P\left(y_i = d|\, x_i^T\right)$$

This means that the set of maximizers (over p_0 and θ) of

$$\begin{aligned} & E\left[\log\left(\pi\left(y_i; p_0\left(\cdot, x^T\right), x^T, \theta\right)\right) w\left(x_i^T\right)\right] \qquad (4) \\ = \; & E\left[\log\left(\int\left\{p_0\left(\alpha, x^T\right)^{y_{i1}}\left(1-p_0\left(\alpha, x^T\right)\right)^{1-y_{i1}}\prod_{t=2}^{T} P\left(y_{it}\middle| x_i^T, y_{it-1};\theta\right)\right\} dG\left(\alpha\middle| x_i^T;\theta\right)\right) w\left(x_i^T\right)\right] \end{aligned}$$

is the set of p_0 and θ such that

$$\pi\left(d; p_0\left(\cdot, x^T\right), x^T, \theta\right) = P\left(d\middle| x^T\right)$$

for all sequences d and for almost all x^T. Hence, as in the minimum distance method, the identified set is the maximand of the objective function (4). In addition, one can obtain the set estimator by those parameter values that come within ϵ_n of the empirical analog of (4).

Maximizing the likelihood over all distributions for the initial conditions as well as over the other parameters of the model was first proposed by Heckman (1981b), who implicitly assumed that the underlying model is identified. This estimator is interesting because it corresponds to the idea that one can "solve" the initial conditions problem by specifying a flexible functional form for $p_0\left(\alpha, x^T\right)$ (see, for example, the discussion in Heckman (1981a, 1981b)). If one interprets the flexible functional form as an implementation of a sieve estimator, then this method can be seen as an application of (4) with the important caveat that the maximum likelihood estimator will, loosely speaking, eventually be in the identified region. To consistently estimate the *whole* identified region, one needs to look at the set of all the parameters values that are close to maximizing the likelihood function.[2]

2.2.3 Linear Programming: the Case for Discrete X

If x_i and α have a discrete distribution then it is possible to derive a different characterization of the identified region for θ, which is more convenient than the minimum distance and maximum likelihood characterizations above. Suppose α has a discrete distribution with a known maximum number of points of support, M. The points of support are denoted by a_m and the probabilities

[2] A recent paper by Chernozhukov, Hahn, and Newey (2004) considers maximum–likelihood estimation of the bounds of a the parameters of a related, but different, panel data discrete choice model.

by ρ_m. For the moment ignore x_i^T. We can then write

$$\begin{aligned}\pi\left(\mathcal{A};p_0\left(\cdot\right),\theta\right) &= \sum_{m=1}^{M}\rho_m\left(p_0\left(a_m\right)\pi\left(\mathcal{A}\right|y_0=1;\theta\right)+\left(1-p_0\left(a_m\right)\right)\pi\left(\mathcal{A}\right|y_0=0;\theta\left.\right)\right) \\ &= \sum_{m=1}^{M}z_m\pi\left(\mathcal{A}\right|y_0=1;\theta\left.\right)+\sum_{m=1}^{M}z_{M+m}\pi\left(\mathcal{A}\right|y_0=0;\theta\left.\right)\end{aligned}$$

where

$$z_m=\rho_m p_0\left(a_m\right)$$

and

$$z_{M+m}=\rho_m\left(1-p_0\left(a_m\right)\right)$$

The sharp identified set Θ, consists of the values of θ for which the equations

$$\begin{aligned}\sum_{m=1}^{M}z_m\pi\left(A\right|y_0=1;\theta\left.\right)+\sum_{m=1}^{M}z_{M+m}\pi\left(A\right|y_0=0;\theta\left.\right) &= P\left(A\right) \qquad \text{for all } A\in\mathcal{A} && (5)\\ \sum_{m=1}^{2M}z_m &= 1 && (6)\\ z_m &\geq 0 && (7)\end{aligned}$$

have a solution for $\left\{z_m\right\}_{m=1}^{2M}$.

Equations (5)–(7) have exactly the same structure as the constraints in a linear programming problem, so checking whether a particular θ belongs to Θ can be done in the same way one checks for a feasible solution in a linear programming problem that has (5)–(7) as the constraints. Specifically, consider the linear programming problem

$$\underset{\{z_m\},\{v_i\}}{\text{maximize}}\sum_i -v_i \qquad (8)$$

$$\begin{aligned}P\left(A\right)-\sum_{m=1}^{M}z_m\pi\left(A\right|y_0=1;\theta\left.\right)-\sum_{m=1}^{M}z_{M+m}\pi\left(A\right|y_0=0;\theta\left.\right) &= v_A \qquad \text{for all } A\in\mathcal{A} && (9)\\ 1-\sum_{m=1}^{2M}z_m &= v_0 && (10)\\ z_m &\geq 0 && (11)\\ v_i &\geq 0 && (12)\end{aligned}$$

This problem clearly has a feasible solution (namely $v_A=P\left(A\right)$ for $A\in\mathcal{A}$, $v_0=1$ and $z_m=0$ for $m=1,...,2M$), and the optimal function value will be 0 if and only if all $v_i=0$, i.e, if a solution exists to (5)–(7). If (5)–(7) do not have a solution, then the maximum function value in (8) is negative. For a given value of θ, one can therefore check whether it belongs to Θ, by solving a linear programming problem and comparing the optimal function value to 0. Alternatively, if one defines $Q(\theta)$ to be

$$Q(\theta) = \max_{\{z_m\},\{v_i\}} \sum_i -v_i \text{ subject to (9)–(12)}$$

then the identified region for θ is the set of maximizers of $Q(\cdot)$. A consistent estimator of the identified region can then be obtained by replacing $P(A)$ in (9) by a consistent estimator, and checking whether, for a give θ, the resulting objective function is within ϵ_n of the maximum value of 0 (or within ϵ_n of the optimal function value).

Provided that x_i is discrete, one can mimic this argument for each value in the support of x_i which will then contribute a set of constraints to the linear programming problem.

2.3 Using single index restrictions

The optimization problems (3) and (4) require optimization over $p_0(\cdot,\cdot)$, which is a function that maps all possible values of x_i and α_i to the interval $[0,1]$. If x is multi-dimensional, then this may be very difficult. Moreover (3) involves the conditional probability of all choice sequences conditional on x_i. In such cases it may be useful to consider other restrictions that can reduce the dimensionality of the problem. For example, suppose that y_{it} is generated by the probit model

$$y_{it} = 1\left\{y_{i,t-1}\gamma + \delta_t + x_i'\beta + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad \text{for} \qquad t = 1, 2, ..., T \tag{13}$$

where ε_{it} is independent of $\{x_i, \alpha_i\}$ and δ_t is a set of time–dummies. Without the time–dummies, and if it is reasonable to assume that (13) has been in effect for a long time before the start of the sample, it might be reasonable to assume that $p_0(\alpha, x_i) = P(y_{i0} = 1|\, x_i, \alpha_i)$ is the stationary distribution of y_{it} given (x_i, α_i) and in that case one would not have an initial conditions problem. See for example the discussion in Heckman (1981b) and Card and Sullivan (1988). With the time–dummies, the process will not be stationary and this approach will not necessarily work. But if the process has been going on "forever" then $p_0(\alpha, x_i) = P(y_{i0} = 1|\, x_i, \alpha_i)$ will be a monotone function of $x_i'\beta$, where the actual functional form is unknown because of the nonstationarity introduced by the δ_t's. The distribution of $(y_{i1}, ..., y_{iT})$ given x_i therefore depends on x_i only through $x_i'\beta$ and β is therefore identified up to scale under appropriate regularity conditions since $p_0(\alpha, x_i)$ depends only on $x_i'\beta + \alpha_i$. This can reduce the dimensionality of the problem.

A monotone index assumption can also be justified if the process has a natural finite starting time where the first observation of y is generated by a model that depends on x_i and α_i only through $x_i'\beta + \alpha_i$.

3 Marginal Effects

In nonlinear models like (2), it is often interesting to estimate marginal effects. Using the ideas developed above, we can construct bounds on these marginal effects. To illustrate this, consider the setup in section 2.2.3 and assume that one wants to explore the difference in period $t+1$ choice probabilities between artificially setting $y_{it} = 0$ and setting $y_{it} = 1$ for an individual with explanatory variables x. This difference would be

$$E\left[\Phi\left(x'\beta+\gamma+\alpha\right)-\Phi\left(x'\beta+\alpha\right)\right]=\sum_m\left(\Phi\left(x'\beta+\gamma+a_m\right)-\Phi\left(x'\beta+a_m\right)\right)P\left(\alpha=a_m|\,x\right)$$

$$=\sum_m\left(\Phi\left(x'\beta+\gamma+a_m\right)-\Phi\left(x'\beta+a_m\right)\right)\left(P\left(\alpha=a_m,y_0=1|\,x\right)+P\left(\alpha=a_m,y_0=0|\,x\right)\right)\quad(14)$$

Note that $P\left(\alpha=a_m,y_0=1|\,x\right)$ and $P\left(\alpha=a_m,y_0=0|\,x\right)$ are exactly the z's in section 2.2.3. In other words, for given values of γ and β in the identified region, we can calculate the upper and lower bounds on the marginal effects by maximizing and minimizing the linear function (14) subject to the linear constraints in section 2.2.3. This can easily be done by linear programming. To find the overall bounds, one can then minimize and maximize these bounds over γ and β in the identified region. Of course, there are many ways to define marginal effects and the specifics of the calculations will depend on which marginal effect is of interest. But it is clear that these marginal effect can be easily constructed using the ideas provided in section 2.2.3 above.

4 Examples

In this section we present a number of examples that illustrate the usefulness of the approach suggested here. The examples are special cases of the probit or logit version of the dynamic discrete choice model

$$y_{it}=1\left\{x_{it}'\beta+y_{i,t-1}\gamma+\alpha_i+\varepsilon_{it}\geq 0\right\}\quad(15)$$

with ε_{it} i.i.d. and $N(0,1)$ or logistically distributed.

Although the examples are motivated by computational simplicity, they are all models for which it is not known whether the parameters of interest are point identified. It is therefore of interest to investigate the identified region for these examples. All of the examples have aggregate explanatory variables x_{it}, and

$$P(\alpha_i = a_j) = \begin{cases} \Phi\left(\frac{a_j + a_{j+1}}{2}\right) & \text{for} \quad a_j = -3.0 \\ \Phi\left(\frac{a_j + a_{j+1}}{2}\right) - \Phi\left(\frac{a_j + a_{j-1}}{2}\right) & \text{for} \quad a_j = -2.8, -2.6, .., 2.8 \\ 1 - \Phi\left(\frac{a_j + a_{j-1}}{2}\right) & \text{for} \quad a_j = 3.0 \end{cases}$$

In words, the true distribution of unobserved heterogeneity is discrete but it closely resembles a standard normal.

While the assumption that the explanatory variable is the same across the individuals makes the calculations much easier, it is also made in order to contrast the matching approach in Honoré and Kyriazidou (2000). If the explanatory variables are independent across individuals and satisfy a support condition, then we know from that paper that the parameters of the model are identified with more than four time–periods. The calculations below will demonstrate that identification can fail with simple violations of this support condition.

We use the linear programming method to compute the identified set in all the examples.

4.1 Only lagged dependent variable

Consider a model with no regressors

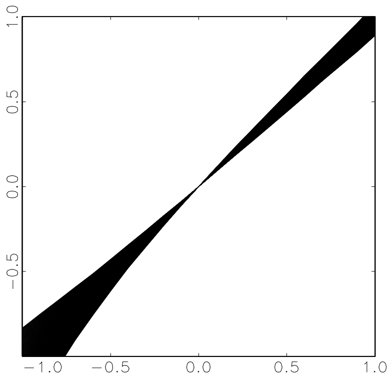
$$y_{it} = 1\{y_{i,t-1}\gamma + \alpha_i + \varepsilon_{it} \geq 0\} \qquad \text{for} \qquad t = 1, 2, ..., T \tag{16}$$

$P(\alpha_i = a_j) = \rho_j$ for $j = 1, ..., 31$, and $P(y_{i0} | \alpha_i) = 0.5$. In calculating the identified region for γ, we assume that it is known that α_i is discrete and that the points of support are $\{-3.0, -2.8, ..., 2.8, 3.0\}$, but that the associated probabilities are unknown. Since ε_{it} is standard normal, this means that the distribution of α_i is extremely flexible over the relevant region. For $T \geq 4$, it is known from Cox (1958) (see also Chamberlain (1985) and Magnac (2000)) that γ would be identified if ε_{it} has a logistic distribution, but to our knowledge, it is not known whether this result carries over to other distributions for ε_{it}.

Using the linear programming techniques developed in section 2.2.3 , we calculate the identified region for γ as a function of the true γ when $T = 3$. The results are presented in Figure 1 where the upper and lower bound on the identified region for γ is plotted as a function of its true value.

It is clear from Figure 1 that γ is not identified when $T = 3$ if $\gamma \neq 0$. On the other hand, the figure suggests that the sign of γ is identified. Lemma 1 of the Appendix shows that this is indeed the case.

Figure 1: Identified region for γ as a function of its true value



For $T = 4$, a graph similar to that in Figure 1 would suggest that γ is point identified. However, a close inspection of the numbers reveal that this is not the case. For example, if $\gamma = 1$, the identified region is $(0.9998, 1.0003)$.

Similar calculations for the case where ε_{it} is logistically distributed yields a graph like that in Figure 1 for $T = 3$, and confirms that γ is point–identified for $T = 4$.

4.2 Lagged dependent variable and time–trend

Of course, many applications do have explanatory variables. If these are individual specific, then the linear programming approach becomes somewhat more cumbersome as each value of x yields constraints of the form (5)–(7). On the other hand, a number of examples include only aggregate variables, such as time trends and time dummies, as explanatory variables. The linear programming technique makes it relatively straightforward to calculate the identified region in cases like this.

As an example, consider the same design as in the previous example, but we include a time trend

$$y_{it} = 1\left\{y_{i,t-1}\gamma + t\beta + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad \text{for} \qquad t = 1, 2, ..., T \tag{17}$$

with ε_{it} i.i.d. standard normal.

Models with time trends are interesting because some of the existing techniques for dealing with models like (15) are based on matching values of x_{it} over different time periods. For example, Honoré and Kyriazidou (2000) show that if $x_{i4} - x_{i3}$ has support in a neighborhood of 0, then (γ, β) in (15) is identified up to scale with $T \geq 4$ even if the distribution of ε_{it} is unknown. The scale is also identified if ε_{it} is logistic. The time trend in (17) is a simple case in which such a matching strategy fails.

Figures 2 and 3 give the identified regions in this case for nine combinations of (γ, β).

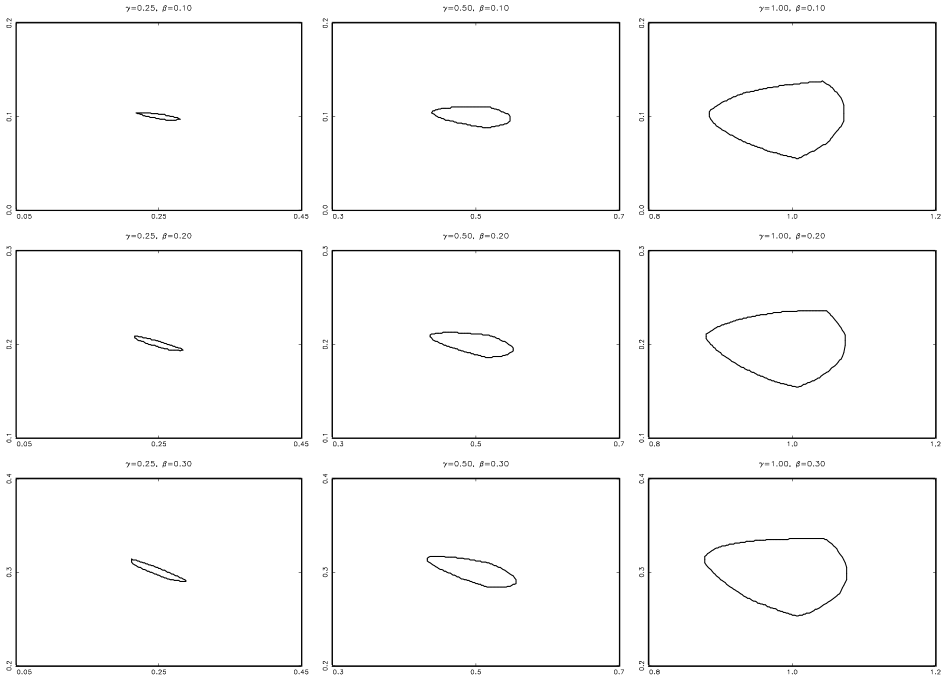


Figure 2: Identified region for (γ, β). $T = 3$

It is not surprising that (γ, β) is not point–identified with $T = 3$ since γ would not be identified even without the time trend. It is interesting that the identified region for (γ, β) is not a singleton when $T = 4$. This suggests that the matching approach in Honoré and Kyriazidou (2000) is essential for obtaining point identification. On the other hand, the size of the identified region suggests that the lack of identification is of little practical consequence.

4.3 Lagged dependent variable and time–dummies

A linear time trend like that in the previous example is very dramatic (and often highly unrealistic) when T is big. In this section we therefore investigate identification when it is replaced by a set of unrestricted time–dummies. Specifically, we consider

$$y_{it} = 1\left\{y_{i,t-1}\gamma + \delta_t + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad \text{for} \qquad t = 1, 2, ..., T \tag{18}$$

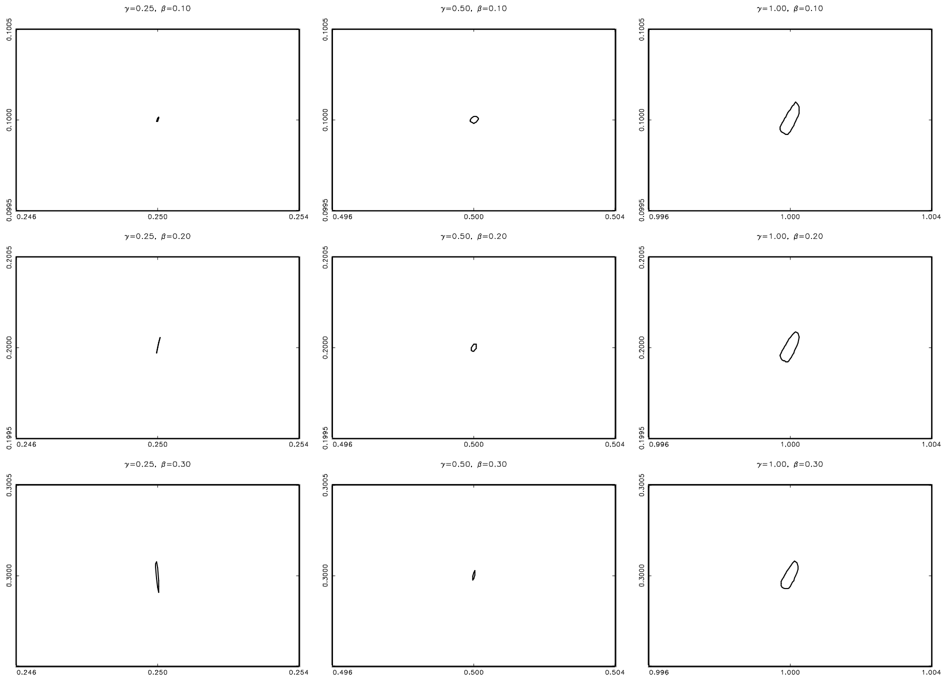


Figure 3: Identified region for (γ, β). $T = 4$

where δ_1 is normalized to 0.[3] Figures 4 and 5 present the identified regions for any pair of the parameters based on $T = 3$ and $T = 4$, respectively, when the true parameters are $\gamma = 1$, $\delta_2 = 0.3$, $\delta_3 = 0.2$ and $\delta_4 = 0.1$.

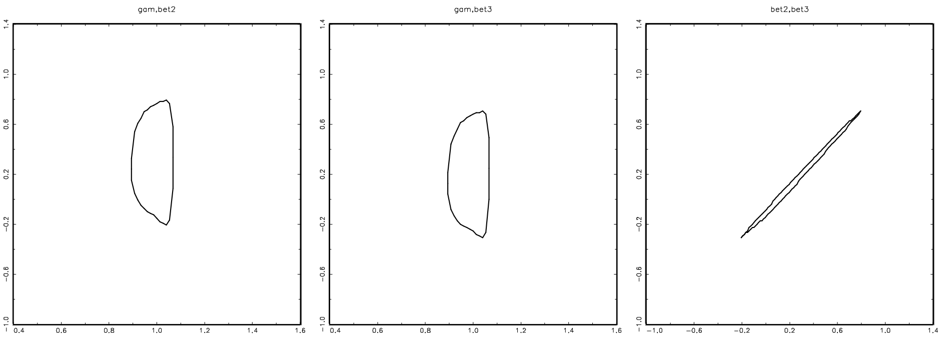


Figure 4: Identified region when $T = 3$, $\gamma = 1$, $\delta_2 = 0.3$ and $\delta_3 = 0.2$

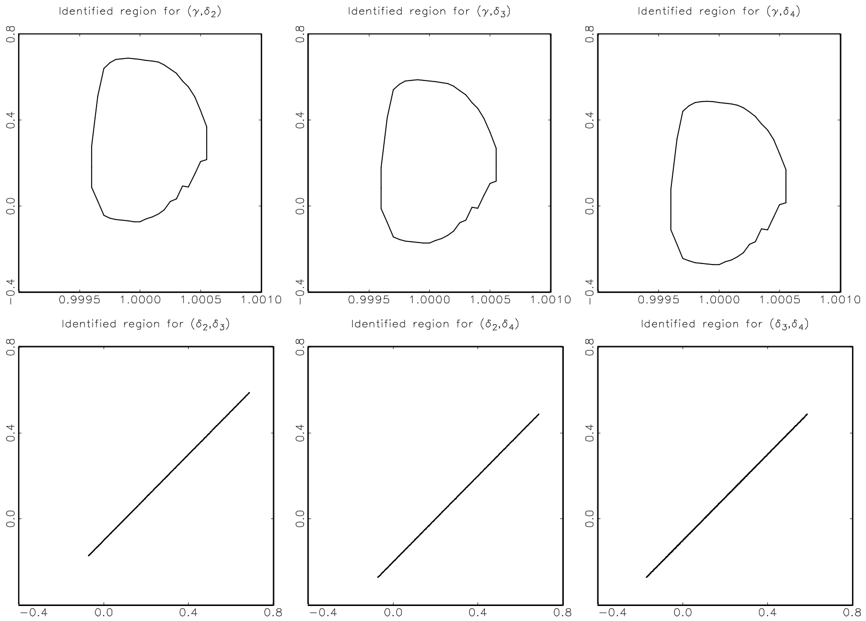


Figure 5: Identified region when $T = 4$, $\gamma = 1$, $\delta_2 = 0.3$, $\delta_3 = 0.2$ and $\delta_4 = 0.1$

The most striking features of Figures 4 and 5 are that it appears that the δ's are identified except for an additive constant and that they are quite poorly identified without such a normalization. The

[3] If the distribution α_i is unrestricted, then a trivial location–normalization would be needed on the δ's. Formally, the distribution of α_i used here is not unrestricted, but it is very flexible, and we therefore impose such a location normalization.

first feature is an artifact of the precision of the figure. An inspection of the actual numbers reveals that what appear to be line–segments in Figure 4 are actually two–dimensional sets with a non–empty interior. The intuition for why it appears that the time–dummies need an additional location normalization is that the unobserved y_{i0} will have a positive effect on all future probabilities. Since the distribution of y_{i0} is unspecified, this would mean that it is difficult to separate the location of this distribution from an additive constant in the δ's. Smaller values of γ would make the effect of the distribution of y_{i0} look less like a constant over time, and one would therefore expect smaller identified regions when γ is smaller. This is confirmed in Figure 6, which presents the results for $\gamma = 0.2$.

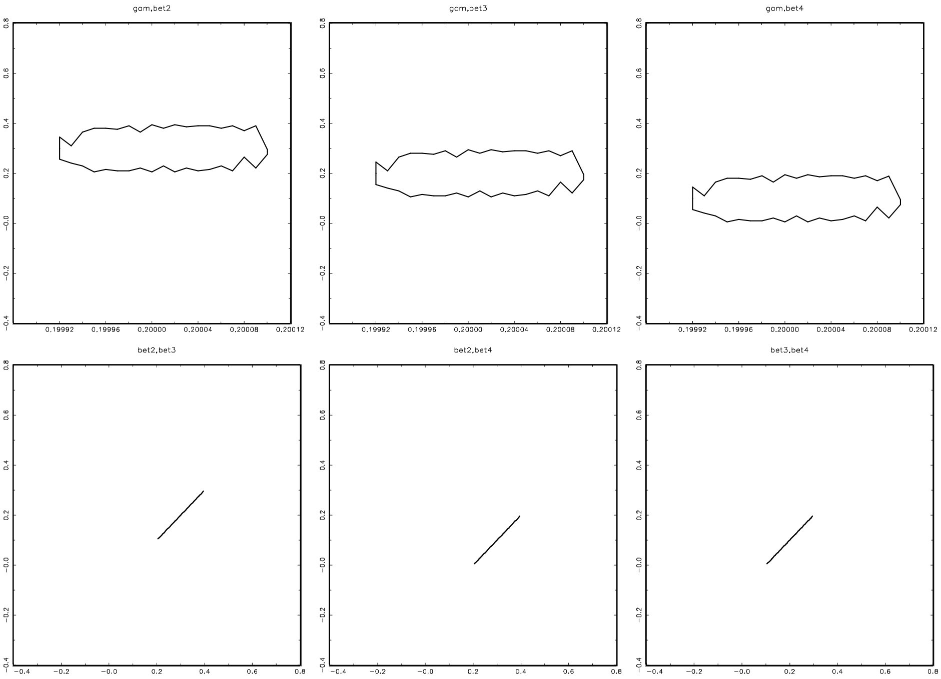


Figure 6: Identified region when $T = 4$, $\gamma = 0.2$, $\delta_2 = 0.3$, $\delta_3 = 0.2$ and $\delta_4 = 0.1$

Overall, the results presented in Figures 1–6 suggest that identification of dynamic discrete choice models relies critically on the ability to match explanatory variables in different time–periods as was done in Honoré and Kyriazidou (2000)[4].

[4] The assumptions on the individual specific effects made here are stronger than the assumptions usually made in the fixed effects literature (see for example Honoré and Kyriazidou (2000)). The nonidentification documented here therefore implies that the corresponding fixed effects models are not identified.

5 Conclusion

This paper examines the question of identification in some nonlinear dynamic panel data models. In particular, we focus on the initial condition problem and its effects on identification of the parameters of interest. This is a classic problem in econometrics that dates back to the work of Heckman ((1978, 1981a, 1981b))). We provide insights that lead to new ways in which identification can be examined and illustrate our approach using the probit version of the dynamic discrete choice model. We give three methods that can be used to construct the identified sets. These methods are constructive in that they can be used, by way of the analogy principle, to obtain consistent estimates of these identified set. In particular, a linear programming method proved to be especially convenient and practical in constructing the identified set when the regressors are discrete.

References

Arellano, M., and R. Carrasco (2003): “Binary Choice Panel Data Models with Predetermined Variables,” *Journal of Econometrics*, 115, 125–157.

Balestra, P., and M. Nerlove (1966): “Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas,” *Econometrica*, 34, 585–612.

Card, D., and D. Sullivan (1988): “Measuring the Effects of Subsidized Training Programs on Movements In and Out of Employment,” *Econometrica*, 56, 497–530.

Chamberlain, G. (1985): “Heterogeneity, Omitted Variable Bias, and Duration Dependence,” in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman, and B. Singer, no. 10 in Econometric Society Monographs series,, pp. 3–38. Cambridge University Press, Cambridge, New York and Sydney.

Chay, K. Y., H. Hoynes, and D. Hyslop (2001): “A Non–Experimental Analysis of True State Dependence in Monthly Welfare Participation Sequences,” University of California, Berkeley.

Chernozhukov, V., J. Hahn, and W. Newey (2004): “Bound Analysis in Panel Models with Correlated Random Effects,” unpublished working paper.

Chintagunta, P., E. Kyriazidou, and J. Perktold (2001): “Panel Data Analysis of Household Brand Choices,” *Journal of Econometrics*, 103(1-2), 111–53.

Contoyannis, P., A. M. Jones, and N. Rice (2002): “Simulation-based Inference in Dynamic Panel Probit Models: an Application to Health,” McMaster University Working paper.

Cox, D. R. (1958): “The Regression Analysis of Binary Sequences,” *Journal of the Royal Statistical Society*, 20, 215–242.

Halliday, T. (2002): “Heterogeneous State Dependence in Health Processes,” Princeton University.

Heckman, J. J. (1978): “Simple Statistical Models for Discrete Panel Data Developed and Applied to Tests of the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence,” *Annales de l’INSEE*, pp. 227–269.

——— (1981a): “Heterogeneity and State Dependence,” *Studies in Labor Markets, S. Rosen (ed).*

——— (1981b): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," *Structural Analysis of Discrete Panel Data with Econometric Applications, C. F. Manski and D. McFadden (eds)*, pp. 179–195.

Honoré, B. E. (2002): "Nonlinear Models with Panel Data," *Portuguese Economic Journal*, 1, 163–179.

Honoré, B. E., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–94.

Magnac, T. (2000): "State Dependence and Unobserved Heterogeneity in Youth Employment Histories," *Economic Journal*, 110, 805–837.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

Wooldridge, J. M. (2000): "The Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," Michigan State University.

6 Appendix

Lemma 1 (1) *Suppose* (y_{i1}, y_{i2}, y_{i3}) *is a random vector such that*

$$P\left(y_{i1}=1\right|\alpha_i)=p_1\left(\alpha_i\right)$$

and

$$P\left(y_{it}=1\right|\alpha_i, y_{i1}, ..., y_{it-1})=F\left(\alpha_i+\gamma y_{it-1}\right), \qquad for \qquad t=2,3$$

where p_1 *is an unknown function taking between 0 and 1 and is an unknown and strictly increasing distribution function. Then the sign of* γ *is identified.*

Proof.

Consider the probabilities

$$P\left(\left(y_{i1}, y_{i2}, y_{i3}\right)=(0,1,0)\right|\alpha_i)=\left(1-p_1\left(\alpha_i\right)\right)\cdot F\left(\alpha_i\right)\cdot\left(1-F\left(\alpha_i+\gamma\right)\right)$$

and

$$P\left(\left.(y_{i1}, y_{i2}, y_{i3}) = (0,0,1)\right|\alpha_i\right) = (1 - p_1(\alpha_i)) \cdot (1 - F(\alpha_i)) \cdot F(\alpha_i)$$

Clearly

$$P\left(\left.(y_{i1}, y_{i2}, y_{i3}) = (0,1,0)\right|\alpha_i\right) \lesseqgtr P\left(\left.(y_{i1}, y_{i2}, y_{i3}) = (0,0,1)\right|\alpha_i\right) \Longleftrightarrow 0 \lesseqgtr \gamma$$

and hence

$$P\left((y_{i1}, y_{i2}, y_{i3}) = (0,1,0)\right) \lesseqgtr P\left((y_{i1}, y_{i2}, y_{i3}) = (0,0,1)\right) \Longleftrightarrow 0 \lesseqgtr \gamma$$

This shows that the sign of γ is identified.