Identification in Discrete Choice Models with Fixed Effects

Edward G. Johnson*

August 2004

Abstract

This paper deals with issues of identification in parametric discrete choice panel data models with fixed effects when the number of time periods is fixed. I show that, under fairly general conditions, the common parameters in a model of discrete probabilities containing a nonparametric "fixed effect" will be non-identified unless the parametric probability distribution obeys a hyperplane restriction for at least some value of the explanatory variables. Alternatively, if such a hyperplane restriction holds with probability zero, the semiparametric information bound is shown to be zero. I use these results to investigate identification in a variety of panel binary choice models. I show how identification and estimation of several classes of known identified models fit into a unifying framework; I show how to check identification in models with additional parameters including models with factor loadings, random trends, and heteroskedasticity; I show

*Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 4120, Washington, DC 20212; johnson.edward@bls.gov. This paper is based on a chapter from my 2004 Ph.D. thesis at Stanford University. I wish to thank the members of my committee: Tom MaCurdy, Edward Vytlacil, Jay Bhattacharya, and Aprajit Mahajan.

that the semiparametric information bound is zero for a panel probit model with fixed T; and I derive a new functional form that is a generalization of logit that allows identification when $T \geq 3$.

1 Introduction

This paper deals with issues of identification in parametric discrete choice panel data models with fixed effects when the number of time periods T is small. Fixed effects models provide a useful way to deal with unobserved heterogeneity that may be correlated with the explanatory variables. The conventional first-difference or within estimators work for linear models, but for non-linear models it is less clear how to proceed. [1] In particular, this is an issue with models of discrete dependent variables, which are nonlinear by their nature. For example, consider a parametric binary choice panel data model, where the outcome y_{it} takes on the values 0 or 1, and is assumed to result from an unobserved latent variable y_{it}^*:

$$y_{it}^* = \alpha_i + x_{it}\beta + \varepsilon_{it}, \quad y_{it} = \mathbf{1}(y_{it}^* > 0). \tag{1.1}$$

In the absence of individual effects, such a model would usually be estimated by maximum likelihood. In a panel setting, it is still possible to perform maximum likelihood estimation, treating the α_i as parameters to be estimated, but the estimates will in general be inconsistent unless both T and N go to infinity. This is the well known "incidental parameters problem" described by Neyman and Scott (1948). If T remains fixed, the common parameters may in many cases not even be identified. Because T is small and fixed in many panel data sets, this is an important issue.

Several identification results for specific models exist in the literature. For the panel data binary choice model, two of the most notable approaches are conditional maximum likelihood estimation (CMLE) for the fixed effects

[1] Honore (2002) gives a good overview of the issues that arise in non-linear panel data models.

logit model (Rasch 1960, Chamberlain 1984), and the panel version of the maximum score estimator (Manski 1987). The CMLE estimator uses the approach suggested by Andersen (1970) of conditioning on a minimal sufficient statistic for the incidental parameters, in order to construct a function that depends only on the common parameters. For the logit functional form, the sum of the outcomes y_{it} is a sufficient statistic for α_i, and the common parameters can be esimated by using the CMLE. Manski (1987) showed that in model (1.1) with $T \geq 2$, β can be consistently estimated up to scale without assuming a functional form for ε_{it} by using a panel version of the maximum score estimator. However, this estimator requires stronger support conditions on the x variables, and it converges at a rate slower than $\sqrt{n}$.

The literature contains few general approaches to checking identification in discrete choice panel models. The approach of conditioning on a sufficient statistic only works for certain special cases. Furthermore, the non-existence of a sufficient statistic does not imply that the model is not identified, nor does it in general rule out $\sqrt{n}$-consistent estimation.[2] There are also very few results in the literature showing which models are *not* identified. The approach of finding semiparametric information bounds can be useful, but even an information bound of zero does not imply that consistent estimation is impossible, as illustrated by the maximum score estimator. It would be useful to have a general approach for showing when a model is not identified.

This paper makes a contribution in that direction. I develop an approach that can often be used to show when identification is possible or impossible in these types of models. It may also suggest new classes of models that are identified, as well as providing some intuition about these questions. The approach applies to models with discrete dependent variables generally, but the main focus here will be on binary choice models. The paper proceeds as follows. Section 2 develops some necessary conditions for identification

[2]Although Magnac (forthcoming) shows that in the case of a standard binary choice latent variable model with two periods, $\sqrt{n}$-consistent estimation *is* ruled out unless the sum $y_{i1} + y_{i2}$ is sufficient for α_i.

in models with discrete outcomes and fixed effects. Section 3 applies these results to examine binary panel data models where the outcomes are conditionally independent across time. Section 4 concludes.

2 Conditions for Identification

Consider a model in which for each individual i, a dependent variable d is observed to take on one of the K discrete outcomes in the set $\Omega = \{\omega_1, ..., \omega_K\}$. We are concerned with the identification of the parameter θ in a parametric model that gives the probabilities of the K outcomes conditional on an observed matrix of explanatory variables $\mathbf{x}$ and an unobserved individual effect α. (In this section the i subscripts are suppressed.)

$$\left\{ \begin{array}{c} \Pr(d = \omega_1 | \mathbf{x}, \alpha) \\ \vdots \\ \Pr(d = \omega_{K-1} | \mathbf{x}, \alpha) \end{array} \right\} = \left\{ \begin{array}{c} p_1(\mathbf{x}, \theta, \alpha) \\ \vdots \\ p_{K-1}(\mathbf{x}, \theta, \alpha) \end{array} \right\} = \mathbf{p}(\mathbf{x}, \theta, \alpha) \tag{2.1}$$

where $p_1(), ..., p_{K-1}()$ are known parametric functions, and $\theta \in \Theta$ is a parameter vector that is common across individuals. We place no restriction on the distribution of α, other than perhaps restricting it to lie in some convex set A; in particular, the distribution α may depend on $\mathbf{x}$. Because probabilities sum to one the function $\mathbf{p}(\mathbf{x}, \theta, \alpha)$ only specifies the first $K - 1$ independent probabilities, omitting outcome ω_K.

In a panel data setting, Ω would consist of the possible values taken by the dependent variable vector $\mathbf{y_i}$. For example, if there are J discrete values taken by y_{it} at each of T time periods $1, ..., T$, then there are J^T possible outcomes for each individual i, so $K = J^T$. The matrix $\mathbf{x}$ would consist of the explanatory variables for each time period $\{x_1, ..., x_T\}$. The model is not limited to this case; for example, it also accommodates models where the set of possible values for the outcome y_{it} is different in different periods, as well as dynamic discrete choice models. In fact, everything in this section applies

even to a non-panel model with K distinct outcomes, although a panel-like structure will usually be necessary for identification.

Assume that the distribution of $\mathbf{x}$ is such that, as the sample becomes large, we can find the reduced form distribution $\mathbf{p}^*(\mathbf{x}) = Pr(d_i = w|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ denotes the support of $\mathbf{x}$. [3] This reduced form function is an average over the different values of α in the population. We wish to investigate for what class of models we can identify the common parameters θ if we know the reduced form distribution, $\mathbf{p}^*(\mathbf{x})$, even when the distribution of α is unknown. First, formally define what is meant here by identification:

Definition 2.1. *Given a model* $\mathbf{p}(\mathbf{x}, \theta, \alpha)$*, we say a particular value* θ_0 *is* **identified** *if there does not exist a value* $\theta' \neq \theta_0$*, along with distributions* $F'(\alpha|\mathbf{x})$ *and* $F(\alpha|\mathbf{x})$*, such that the reduced form distributions* $\mathbf{p}^*(\mathbf{x})$ *generated by* $\{\theta_0, F(\alpha|\mathbf{x})\}$ *and* $\{\theta', F'(\alpha|\mathbf{x})\}$ *are identical. If there* does *exist such a value and distributions, then we say that* θ_0 *is non-identified relative to* θ'*.*

Intuitively, a value of θ_0 is identified if there is no other value that could have produced the same reduced form. Note that the definition requires this to hold for *any* possible distribution $F(\alpha|\mathbf{x})$ over the set A.

The reduced form distribution $\mathbf{p}^*(\mathbf{x})$ is a vector-valued function in $K-1$ dimensional space, $\mathbf{p}^* : \mathcal{X} \mapsto \mathbb{R}^{K-1}$. This distribution depends on both θ and on the distribution of individual effects $F(\alpha|\mathbf{x})$. Since the function expresses a vector of probabilities, it can only take values such that each element is positive and the sum of the elements does not exceed one.

As α varies (holding $\mathbf{x}$ and θ fixed), the function $\mathbf{p}(\mathbf{x}, \theta, \alpha)$ defines a set of points in $\mathbb{R}^{K-1}$.

$$C_{x\theta} = \{q \in \mathbb{R}^{K-1} : \exists \alpha \ \in A \text{ such that } q = \mathbf{p}(\mathbf{x}, \theta, \alpha)\}$$

If $\mathbf{p}()$ is continuous, then $C_{x\theta}$ will trace out a continuous curve in $K-1$ dimensional space, showing how the probabilities of the outcomes change as α changes.

[3] More precisely, assume we can find a function $\mathbf{p}^*(\mathbf{x})$ that equals $Pr(d_i = w|\mathbf{x})$ almost surely.

The reduced form distribution $\mathbf{p}^*(\mathbf{x})$ is given by the expected value of $\mathbf{p}(\mathbf{x}, \theta, \alpha)$, taken over the distribution of α conditional on $\mathbf{x}$:

$$\mathbf{p}^*(\mathbf{x}) \ = \int_{\alpha \in A} \mathbf{p}(\mathbf{x}, \theta, \alpha) dF(\alpha|\mathbf{x}) d\alpha$$

Therefore, the reduced form distribution at $\mathbf{x}$ is a weighted average of points in the set $C_{x\theta}$, weighted according to $F(\alpha|\mathbf{x})$. For a given $\mathbf{x}$, the set of values that $\mathbf{p}^*(\mathbf{x})$ could possibly take is equal to the convex hull of $C_{x\theta}$. Denote this convex hull as $H_{x\theta}$. With no restriction on $F(\alpha|\mathbf{x})$, each point $\mathbf{p}^*(\mathbf{x})$ could potentially lie anywhere in the set $H_{x\theta}$. In other words, the only restriction that θ places on the observables is that for each value of $\mathbf{x}$, $\mathbf{p}^*(\mathbf{x})$ must lie somewhere in the set $H_{x\theta}$. If an observed outcome $\mathbf{p}^*(\mathbf{x})$ lies in two different convex hulls generated by different values of θ, then identification is not possible. Identification is only possible if the sets $H_{x\theta}$ for different values of θ don't overlap.

More formally, if

$$\exists\, \theta' \neq \theta_0 \text{ such that } \forall \mathbf{x} \in \mathcal{X},\ H_{x\theta_0} \cap H_{x\theta'} \neq \emptyset$$

then θ_0 is not identified relative to θ'.

Note that the distributions $F(\alpha|\mathbf{x})$ are not identified. In fact, according to Caratheodory's theorem, every point in the convex hull of a set in $\mathbb{R}^{K-1}$ can be generated by a convex combination of at most K points from the set. Therefore, every possible reduced form function $\mathbf{p}^*(\mathbf{x})$ can be generated by some set of K-point discrete distributions $F(\alpha|\mathbf{x})$, and without further restrictions we can never distinguish between those distributions and other distributions that may have more than K points of support.

I now show that a necessary condition for identification under general assumptions is that the set of values $C_{x\theta}$ that can be taken by $\mathbf{p}(\mathbf{x}, \theta, \alpha)$ as α varies must lie entirely in a $K-2$ dimensional hyperplane, for at least one value of $\mathbf{x} \in \mathcal{X}$.

Theorem 2.2. *Assume the following assumptions hold for a model such as equation (2.1), and let θ_0 be some particular value of θ:*

1. *The $K-1$ dimensional function $\mathbf{p}(\mathbf{x}, \theta, \alpha)$ is continuous in $\mathbf{x}$, θ and α, at $\theta = \theta_0$.*

2. *The set A of values that may be taken by α is convex.*

3. *The limits $\lim_{\alpha \to \inf(A)} \mathbf{p}(\mathbf{x}, \theta_0, \alpha)$ and $\lim_{\alpha \to \sup(A)} \mathbf{p}(\mathbf{x}, \theta_0, \alpha)$ exist and are continuous (or constant) functions of $\mathbf{x}$.*

4. *For any sequence $\theta_n \to \theta_0$, with $\theta_n \neq \theta_0$, $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ converges uniformly in $\mathbf{x}$ and α to $\mathbf{p}(\mathbf{x}, \theta_0, \alpha)$, and at least one such sequence exists.*

5. *The support of $\mathbf{x}$ is confined to a compact set $\mathcal{X}$.*

Then a necessary condition for θ_0 to be identified is that the set of points $C_{x\theta_0} = \{q \in \mathbb{R}^{K-1} : \exists \alpha \in A,$ such that $q = \mathbf{p}(\mathbf{x}, \theta_0, \alpha)\}$ is contained within some $K-2$ dimensional hyperplane for some value of $\mathbf{x} \in \mathcal{X}$. In other words, for some $\mathbf{x}$ there must exist constants $c_0, ..., c_{K-1}$, not all zero, such that

$$\forall \alpha \in A, \ \sum_{j=1}^{K-1} c_j p_j(\mathbf{x}, \theta_0, \alpha) = c_0. \tag{2.2}$$

PROOF: Assume that the necessary condition is not met, so that the set of points $C_{x\theta_0}$ never lies in a hyperplane for any value of $\mathbf{x} \in \mathcal{X}$. Then the convex hull $H_{x\theta_0}$ must always have some $K-1$ dimensional volume. Define the "thickness" of the convex hull $H_{x\theta_0}$ to be the supremum of the radii of all the $K-1$ dimensional open balls that lie within $H_{x\theta_0}$. The assumption that $C_{x\theta_0}$ does not lie in a hyperplane guarantees that this thickness is never zero. Because of assumptions 1 and 3, the boundary of the convex hull will move in a continuous fashion as $\mathbf{x}$ changes. This implies that the thickness of $H_{x\theta_0}$ will also be a continuous function of $\mathbf{x}$. Because $\mathcal{X}$ is compact, there is some minimum thickness. Therefore there exists some $\epsilon > 0$ such that, for every $\mathbf{x}$, there is an open ball of radius ϵ entirely within $H_{x\theta_0}$.

For each value of $\mathbf{x}$, let $a(\mathbf{x})$ denote a $K-1$ dimensional point that is the center of some open ball with radius ϵ within $H_x\theta_0$. By Caratheodory's

theorem, we can always write $a(\mathbf{x})$ as a convex combination of K points from the set $\mathbf{p}(\mathbf{x}, \theta, \alpha)$:

$$a(\mathbf{x}) = \sum_{m=1}^{K} w_m(\mathbf{x})\mathbf{p}(\mathbf{x}, \theta, \alpha_m(\mathbf{x})), \quad \sum_{m=1}^{K} w_m(\mathbf{x}) = 1, \quad w_1(\mathbf{x}), ..., w_K(\mathbf{x}) \geq 0$$

where the points $\alpha_m(\mathbf{x})$ and the weights $w_m(\mathbf{x})$ may depend on $\mathbf{x}$, .

By assumption 4 there exists some $\theta' \neq \theta_0$ such that $\forall \mathbf{x} \in \mathcal{X}, \alpha \in A, \ \|(\mathbf{p}(\mathbf{x}, \theta', \alpha) - \mathbf{p}(\mathbf{x}, \theta_0, \alpha)\| < \epsilon$. For each value of $\mathbf{x}$ define the point $a'(\mathbf{x}) = \sum_{m=1}^{K} w_m \mathbf{p}(\mathbf{x}, \theta', \alpha_m)$. Using the triangle inequality,

$$\begin{aligned} \|a'(\mathbf{x}) - a(\mathbf{x})\| &= \|\sum_{m=1}^{K} w_m(\mathbf{p}(\mathbf{x}, \theta', \alpha_m) - \mathbf{p}(\mathbf{x}, \theta_0, \alpha_m))\| \\ &\leq \sum_{m=1}^{K} w_m \|\mathbf{p}(\mathbf{x}, \theta', \alpha_m) - \mathbf{p}(\mathbf{x}, \theta_0, \alpha_m)\| < \epsilon \end{aligned}$$

Therefore $a'(\mathbf{x})$ lies in the convex hull $H_{x\theta_0}$ for each value of $\mathbf{x}$. So the reduced form represented by the point $a'(\mathbf{x})$ could have been generated by θ' or θ_0, and θ_0 is not identified. □

The reason that the theorem is true should be easy to see: if the set of points defined by varying α does not lie in a hyperplane, then the convex hull of those points contains some K-1 dimensional volume. And if the boundaries of these sets move continuously with θ, then they must overlap. The assumptions in the theorem serve to rule out cases where the convex hulls get arbitrarily thin, or where the boundaries move arbitrarily quickly, which might otherwise allow identification. All the assumptions on the functional form of $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ are met for most functional forms encountered in practice, including panel logit and panel probit models. Some of the assumptions of the theorem could be relaxed. In particular, the set A of α could be allowed to depend on $\mathbf{x}$, as long as the boundaries of this set move continuously as $\mathbf{x}$ changes. It would also be possible to have more than one individual effect, i.e. α could be a vector. The assumptions would have to be suitably modified

to ensure that the convex hulls do not become arbitrarily thin and that their boundaries move continuously.

Notice that the theorem rules out even up-to-scale parameter identification, since we can use a sequence θ_n converging to θ_0 from any direction.

The necessary condition in Theorem 2.2 requires that the hyperplane restriction holds for some value of $\mathbf{x} \in \mathcal{X}$. If a hyperplane restriction holds only for a set values of $\mathbf{x}$ which occur with zero probability, we cannot rule out identification using Theorem 2.2, but we may be able to rule out $\sqrt{n}$-consistent estimation because the semiparametric information bound will be zero, as shown in the following theorem. The following theorem also allows us to handle cases where the support $\mathcal{X}$ is not compact.

Theorem 2.3. *Assume there exists a sequence of compact sets $\mathcal{X}_n \in \mathcal{X}$ such that $\Pr(\mathbf{x} \in \mathcal{X}_n) \to 1$ and such that no hyperplane restriction (2.2) holds for any $\mathbf{x} \in \mathcal{X}_n$. Let assumptions 1, 2, 3, and 4 from theorem 2.2 hold. In addition assume:*

1. *$\mathbf{p}(\mathbf{x}, \theta_0, \alpha)$ is continuously differentiable in θ in a neighborhood of θ_0.*
2. *There exists some functions $q_{jk}(\mathbf{x})$ such that for some $\alpha' \in A$,*

$$p_k^{-1/2}(\mathbf{x}, \theta_0, \alpha')\partial p_k(\mathbf{x}, \theta_0, \alpha')/\partial\theta_j < q_{jk}(\mathbf{x})$$

for all θ in a neighborhood of θ_0, and $E(q_{jk}(\mathbf{x})^2) < \infty$.

Then the semiparametric information bound for θ_0 is zero.

A proof is provided in the appendix. The intuition behind this result is simple: θ_0 is not identified by observations on a compact set $\mathcal{X}_n$, so no "information" is provided by observations inside such a set. If the probability of being outside such a set is arbitrarily small, then the expected "information" (i.e. the square of the score) will also be arbitrarily small, as long as the score is bounded.

In many cases it may be difficult to show whether the function $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ obeys a hyperplane restriction. The following lemmas can be useful. (See the next section for examples).

Lemma 2.4. *Assume that there exists some sequence $\alpha_n \in A$ such that $\lim_{n\to\infty} \mathbf{p}(\mathbf{x}, \theta_0, \alpha_n) = (0, ..., 0)$. (In other words, at the limit the omitted outcome K has probability 1).*

Then the hyperplane restriction (2.2) may hold only with $c_0 = 0$. In other words, it may only hold if the elements of $\mathbf{p}(\mathbf{x}, \theta_0, \alpha)$, considered as functions of α, are linearly dependent.

PROOF: Since by assumption the hyperplane must come arbitrarily close to the origin, it must pass through the origin, and there must exist some vector $\mathbf{c}$ normal to the hyperplane. □

Lemma 2.5. *Call two probabilities p_j and p_k, which are components of $\mathbf{p}(\mathbf{x}, \theta_0, \alpha)$, **limit-incommensurate** if one of the limits*

$$lim_{\alpha\to \inf(A)} \frac{p_j(\mathbf{x}, \theta_0, \alpha_i)}{p_k(\mathbf{x}, \theta_0, \alpha)}$$

$$lim_{\alpha\to \sup(A)} \frac{p_j(\mathbf{x}, \theta_0, \alpha_i)}{p_k(\mathbf{x}, \theta_0, \alpha)}$$

is either 0 or ∞.

Then if all pairs elements from $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ are limit-incommensurate, the function $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ is not linearly dependent

PROOF: Linear dependence requires that there exists some vector $\mathbf{c} \neq \mathbf{0}$ such that for all values of $\alpha_i \in A$, $\sum_{j=1}^{K-1} c_j p_j(x, \theta_0, \alpha_i) = 0$. Choose k such that $c_k \neq 0$, and $lim \frac{p_j(x,\theta_0,\alpha_i)}{p_k(x,\theta_0,\alpha_i)} = 0$ for all j such that $c_j \neq 0$. Rearrange to get: $\sum_{j\neq k} c_j \frac{p_j(x,\theta_0,\alpha_i)}{p_k(x,\theta_0,\alpha_i)} = -c_k$. If all probabilities are limit-incommensurate, the left hand side will converge to 0, a contradiction. □

Thus, if for all values of $\mathbf{x_i} \in \mathcal{X}$, all pairs of probabilities in $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ are limit-incommensurate, and if $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ comes arbitrarily close to the origin, then identification is impossible.

Comments

Note that the hyperplane restriction (2.2) in Theorem 2.2 is necessary but not sufficient for identification of θ_0. In any given class of models there may be additional requirements on the specific functional form, or on the support of $\mathbf{x}$. For example, in a standard linear panel data model, an intercept coefficient will not be identified. In addition, there will be requirements on the support of $\mathbf{x}$ to rule out perfect multicolinearity. However, these types of considerations are often simple to understand in any specific case, and will not be the focus of discussion in this paper.

If the necessary condition of the theorem holds and $\mathbf{p}(\mathbf{x}, \theta_n, \alpha)$ lies in a hyperplane, estimation can often be accomplished by using a method of moments estimator that imposes this hyperplane restriction. See the next section for examples.

3 Application: Binary Choice Models

In this section I apply the theorem from the last section to the case of panel binary choice models, with $y_1, ..., y_T$ independent of each other conditional on $\mathbf{x_i}$ and α_i.

The section proceeds as follows: I first lay out some notation and maintained assumptions. Next I review some previous literature on these models. I then use a simple graphical example of fixed effect probit and logit models to illustrate and give intuition, and I show some results for the probit model when $T \neq 2$. Next I discuss some classes of models that are known to be identified and show how identification and estimation of these models fit into a unifying framework. I then apply the approach to check identification in models with additional individual specific parameters, including factor loadings, random trends, and heteroskedasticity. Finally, I derive a new functional form that is a generalization of the logit and for which identification is possible when $T \geq 3$.

3.1 Notation and assumptions

The models discussed in this section have the following form:

$$\Pr(y_{it} = 1|x, \alpha_i) = G(x_{it}, \alpha_i, \theta), \tag{3.1}$$

where each individual i is observed for T periods, $t = 1, ..., T$, y_{it} takes on the values 0 or 1, and $G()$ is a known parametric function. This model can always be written (in an infinite number of different ways) as a latent variable model:

$$y^*_{it} = g(\mathbf{x_i}, \alpha_i, \theta) + \varepsilon_{it},\ y_{it} = \mathbf{1}(y^*_{it} > 0). \tag{3.2}$$

A more restrictive form that appears often in the literature is the linear latent variable model:

$$y^*_{it} = \alpha_i + x'_{it}\theta + \varepsilon_{it},\ y_{it} = \mathbf{1}(y^*_{it} > 0) \tag{3.3}$$

Hereafter I maintain the following two assumptions:

EXOGENEITY: $\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = \Pr(y_{it} = 1|x_{it}, \alpha_i)$

INDEPENDENCE: $y_{i1}, ..., y_{it}$ are independent conditional on $\mathbf{x_i}, \alpha_i$.

To simplify notation in this section, write the function $\mathbf{p}(\mathbf{x}, \theta, \alpha)$ from the last section as $\mathbf{p}_{x\theta}(\alpha)$. This function gives the $2^T - 1$ vector of probabilities of the possible outcomes $d = (y_1, y_2, ..., y_T)$. Assume that the all-zeros outcome $(0, 0, ..., 0)$ is omitted. It will be understood that $\mathbf{p}_{x\theta}(\alpha)$ is a function of α_i, holding the values of $\mathbf{x_i}$ and θ fixed. Denote the probability of a given outcome as $p_{y_1 y_2 \cdots y_T}$. For example, if T=2, then $\mathbf{p}_{x\theta}(\alpha)$ is a function that gives the three probabilities $\{p_{01}, p_{10}, p_{11}\}$. Let $p_j(\alpha_i)$ denote some element of this vector. Because of the independence assumption, the elements of this function are just products of the individual probabilities. Let S_j denote the set of values t for which $y_{it} = 1$ in outcome j. For example, if outcome j is (1011), then $S_j = \{1, 3, 4\}$, and

$$p_j(\alpha_i) \equiv p_{1011} = \prod_{t \in S_j} G(x_{it}, \alpha_i, \theta) \prod_{t \notin S_j} (1 - G(x_{it}, \alpha_i, \theta)).$$

For various functional forms I will check if (or when) the function $\mathbf{p}_{x\theta}(\alpha)$ obeys a hyperplane restriction necessary for identification according to Theorem 2.2. To ensure that the assumptions of the theorem are met, all the functional forms I consider will satisfy the following assumptions on the $G()$ function.

> CONTINUITY: $G(x, \theta_n, \alpha_i)$ is continuous in x, θ and α_i, and for any sequence $\theta_n \to \theta_0$, $G(x, \theta_n, \alpha_i)$ converges uniformly to $G(x, \theta_0, \alpha_i)$.

I make the following assumption on the support of $\mathbf{x_i}$:

> COMPACTNESS: The support of $\mathbf{x_i}$ is confined to a compact set $\mathcal{X}$.

This is the only assumption that is not standard in models of this type. I make it here because I'm interested in what identification can be achieved without strong conditions on the support of $\mathbf{x_i}$. Furthermore, one might doubt the practical usefulness of an estimator that only gives identification "at infinity." Without this assumption it is difficult to make general statements ruling out cases where θ is identified at the limits of the support (although Theorem 2.3 can still be used to rule out $\sqrt{n}$ convergence). In many specific cases, the theorem could be adapted to rule out identification even when $\mathbf{x_i}$ is unbounded.

3.2 Previous literature

I now discuss in more detail two of the most notable results for models of these types: the conditional maximum likelihood estimator (CMLE) for the fixed effects logit model, and the panel maximum score estimator.

The fixed effects logit estimator uses the approach suggested by Andersen (1970) of conditioning on a minimal sufficient statistic for the incidental parameters, in order to construct a function that depends only on the other parameters. It is well known that for the panel logit model, the required sufficient statistic exists, and the common parameters can be identified by

using a conditional maximum likelihood estimator. The panel logit model is given by:

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = \frac{e^{\alpha_i + x'_{it}\beta}}{1 + e^{\alpha_i + x'_{it}\beta}}$$

In this case the sum of the number of '1' outcomes for an individual is a sufficient statistic for α_i. Conditioning on the number of '1' outcomes leads to an expression that does not depend on α_i. For example, if T=2, then we can write:

$$\Pr(y_{i0} = 1, y_{i1} = 0 \mid y_{i0} + y_{i1} = 1, \mathbf{x_i}, \alpha_i) = \frac{1}{1 + e^{(x_{i2} - x_{i1})'\beta}}$$

Maximizing this likelihood using the subset of observations where $y_{i0}+y_{i1} = 1$ yields consistent estimates of β (with the exception of an intercept term, which is not identified). The CMLE is somewhat analogous to a linear first-differences estimator, in that only the differences in the values of x over time are used. However instead of using simple differences in y, we use the probabilities of observing (1,0) versus (0,1).

Manski (1987) uses a clever insight to show that in a latent variable model such as 3.3 with a linear index $g() = \alpha_i + x_{it}\beta$ and $T \geq 2$, β can be identified up to scale without assuming a functional form for ε_{it} by using a panel version of Manski's maximum score (MS) estimator. Identification is accomplished by conditioning on the cases where $y_{i1} + y_{i2} = 1$, just as in the fixed effects logit estimator. The basic insight is that when $y_{i1} + y_{i2} = 1$, the probability of observing (0,1) will be the same as the probability of observing (1,0) if and only if $x_{i1}\beta = x_{i2}\beta$. Therefore if we can observe the set of values of x for which (1,0) and (0,1) are equally likely, we can identify the effects of one x variable relative to another.

One might think that with the panel maximum score estimator, the identification question has been answered, since it allows identification with only weak assumptions on the error terms. However, the maximum score identification result is weaker than the CMLE result in three important ways. First, the MS estimator identifies the parameters only up to scale—that is,

we can only observe the effects relative to each other. [4] Second, the MS estimator requires that $\mathbf{x}$ have support in some region where $x_{i1}\beta = x_{i2}\beta$, with $x_{i1} \neq x_{i2}$. There are common situations where this condition may not be met. For example, the support condition may fail to hold if only one of the variables changes for an individual, if the changes are always all in the same direction, or if all the x variables are discrete. Flexible functional forms, such as those containing several interaction terms, might also present a problem. By contrast, the fixed effects logit approach only requires that the x variables differ over time periods, and that these differences have full rank. Finally, the MS estimator converges at a rate slower than $\sqrt{n}$. This is makes intuitive sense, because given the weak assumptions in the model, as n gets large only observations near the region where $x_{i1}\beta = x_{i2}\beta$ are informative about β.

Other researchers have extended one or both of these approaches to find consistent estimators in related settings. Chamberlain (1984) generalizes the conditioning approach to the multinomial logit model. Honore and Kyriazidou (2000) use a similar approach to show how to estimate a dynamic logit model with serial dependence when the panel is at least four time periods long. Johnson (2004) applies a similar method to a panel ordered logit model. Other researchers have developed new approaches that depend on strong restrictions on the joint distribution of $\mathbf{x_i}$ and α_i to gain identification (for example Honore and Lewbel 2002, Altonji and Matzkin 2001).

[4] It is well known that all standard discrete choice models, such as probit or logit, involve a normalization on the variance of the error term. However, general statements about parameters being identified only "up to scale" or "up to location" can be misleading. Given exogeneity assumptions, a standard parametric probit or logit model, even though it contains a normalization, allows predictions about the marginal effects on outcomes of the changes in explanatory variables. In effect, the scale of the parameters is identified relative to all the omitted factors that make up the error term. Such predictions are impossible from methods such as MS that only find the scale of parameters relative to each other. They do not yield a model of probabilities such as (3.1), and cannot be used for prediction. In fact, the estimates from the MS model place no restrictions at all on the magnitudes of the effects of the x variables on the probabilities.

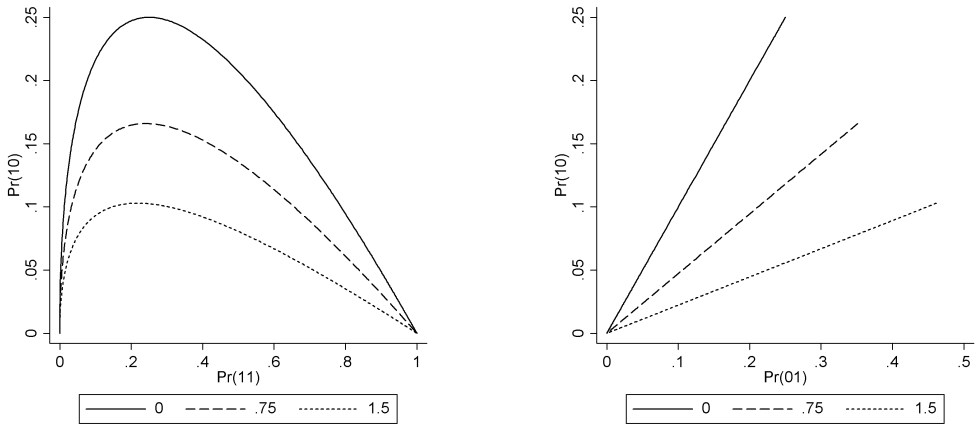

Figure 1: Logit $\mathbf{p}_{x\theta}(\alpha)$ for three different values of $(x_{i2} - x_{i1})'\beta$

3.3 Illustration: logit vs. probit

In applying the theorem in Section 2, it will be helpful to start with a simple graphical example. Consider the fixed effects logit model when T=2 and the index is linear:

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = \frac{e^{\alpha_i + x_{it}'\beta}}{1 + e^{\alpha_i + x_{it}'\beta}} \tag{3.4}$$

With this specification β can be consistently estimated using the CMLE. According to Theorem 2.2, this means $\mathbf{p}_{x\theta}(\alpha)$ must obey a hyperplane restriction for some value of $\mathbf{x_i}$. In fact $\mathbf{p}_{x\theta}(\alpha)$ is linearly dependent for *every* value of $\mathbf{x_i}$, because for any value of $\mathbf{x_i}$, the ratio

$$\frac{p_{01}}{p_{10}} = e^{(x_{i2} - x_{i1})'\beta}$$

does not depend on α_i.

The situation is illustrated graphically in Figure 1. The figure traces out two views of the locus of points in 3-dimensional space representing the values $\mathbf{p}_{x\theta}(\alpha)$ may take on, or in other words the combinations of p_{01},p_{10}, and p_{11} that are possible, as α_i varies from $-\infty$ to ∞, for 3 different values of $(x_{i2} - x_{i1})'\beta$. The observed reduced form $\mathbf{p}^*(\mathbf{x})$ will be some point in the convex hull of the set of points defined by $\mathbf{p}_{x\theta}(\alpha)$.

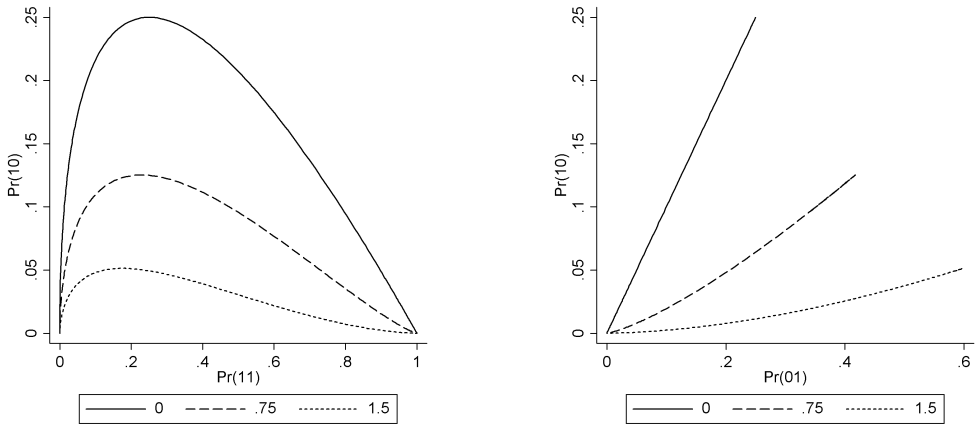

Figure 2: Probit $\mathbf{p}_{x\theta}(\alpha)$ for three different values of $(x_{i2} - x_{i1})'\beta$

For each value of $(x_{i2} - x_{i1})'\beta$, the function lies entirely within a plane, as can be seen in the "side view" graph on the right. The top curve in each figure shows the case where $(x_{i2} - x_{i1})'\beta = 0$. In this case, the points for both functions lie on the plane where $p_{01} = p_{10}$. [5] The lower curves show the values p_{01} and p_{10} can take when $(x_{i2} - x_{i1})'\beta = .75$ or $(x_{i2} - x_{i1})'\beta = 1.5$. Conditional on $(x_{i2} - x_{i1})$, different values of β partition the set of observable reduced forms into non-overlapping hyperplanes. By observing which hyperplane the reduced form falls in, we can determine β. The only requirement on the support of $\mathbf{x_i}$ is the standard assumption that $(x_{i2} - x_{i1})$ may not exhibit perfect multicolinearity.

By contrast, for the probit specification, $\mathbf{p}_{x\theta}(\alpha)$ generally does not lie in a hyperplane. Figure 2 graphically depicts the $\mathbf{p}_{x\theta}(\alpha)$ function for the fixed-effect probit specification:

$$\Pr(y_{it} = 1) = \Phi(\alpha_i + x'_{it}\beta) \tag{3.5}$$

As before, the three curves depict $\mathbf{p}_{x\theta}(\alpha)$ for different values $(x_{i2} - x_{i1})'\beta$. The lower two curves on the right side of Figure 2 are curved, so the function

[5] The scale of the axes makes the curve appear to be steeper than its true 45 degrees.

$\mathbf{p}_{x\theta}(\alpha)$ does not lie within a hyperplane for these values of $(x_{i2}-x_{i1})'\beta$. [6] The only value of $\mathbf{x_i}$ for which $\mathbf{p}_{x\theta}(\alpha)$ does lie in a plane is when $(x_{i2}-x_{i1})'\beta = 0$, which makes $p_{01} = p_{10}$. If the support of $\mathbf{x_i}$ includes values such in a neighborhood such that $p_{01} = p_{10}$ for some $x_{i1} \neq x_{i2}$, then it is possible to identify θ up to scale. This is the principle behind the panel maximum score estimator. If the support of $\mathbf{x_i}$ does not include points where $(x_{i2}-x_{i1})'\beta = 0$, then identification of β is impossible (even up to scale). [7] Even if the support does contain such points, if the distribution of any element of $\mathbf{x_i}$ is continuous such that these points have probability zero, then Theorem 2.3 applies, the semiparametric information bound is zero, and $\sqrt{n}$-consistent estimation is impossible.

Probit when $T \geq 3$

When $T = 2$ it is easy to see that the panel probit model (3.5) does not obey the hyperplane restriction by examining the graph in Figure 2. For values of T larger than 2, we can use Lemma 2.5. Consider a generalized panel probit model with a possibly non-linear index:

$$\Pr(y_{it} = 1) = \Phi(\alpha_i + g(x_{it}, \theta))$$

Let m_j denote the number of 1's in some outcome j, and let S_j denote the set of values of t for which $y_{it} = 1$ in outcome j. Let $g(x_{it})$ denote the function $g(x_{it}, \theta)$. Consider the probabilities for two outcomes j and k:

$$\lim_{\alpha_i \to \infty} \frac{p_j}{p_k} = \lim_{\alpha_i \to \infty} \frac{\prod_{t \notin S_j} \Phi(-\alpha_i - g(x_{it}))}{\prod_{s \notin S_k} \Phi(-\alpha_i - g(x_{is}))}$$

[6] Obviously the functions do not curve very much, meaning that the convex hulls involved are quite "thin." This suggests that even though the parameters are not formally identified in the panel probit model, it would be theoretically possible to identify fairly tight bounds on them. This is not surprising, since the probit specification is quite close to logit, and the logit specification is identified.

[7] Note, however, that the value of $\beta_0 = 0$ *is* identified. This is typical—if the probability of the outcome does not depend on $\mathbf{x_i}$ at all, we can detect that fact.

We can find the value of this limit by observing that the limit

$$\lim_{x\to\infty} \frac{\frac{1}{2x\sqrt{2\pi}}e^{-x^2/2}}{\Phi(-x)} = 1$$

Therefore we can replace each $\Phi(-\alpha_i - g(x_{it}))$ term in the limit (3.3) with $e^{-(\alpha_i+g(x_{it}))^2/2}/(\alpha_i + g(x_{it}))$, which leads to:

$$\begin{aligned}\lim_{\alpha_i\to\infty} \frac{p_j}{p_k} &= \lim_{\alpha_i\to\infty} \frac{\prod_{t\notin S_j} e^{-\frac{1}{2}(\alpha_i+g(x_{it}))^2}}{\prod_{s\notin S_k} e^{-\frac{1}{2}(\alpha_i+g(x_{is}))^2}} \\ &= \lim_{\alpha_i\to\infty} \frac{e^{\frac{1}{2}(-m_j\alpha_i^2-2\alpha_i\sum_{t\notin S_j} g(x_{it})-\sum_{t\notin S_j} g(x_{it})^2)}}{e^{\frac{1}{2}(-m_k\alpha_i^2-2\alpha_i\sum_{s\notin S_k} g(x_{is})-\sum_{s\notin S_k} g(x_{is})^2)}}\end{aligned}$$

Clearly the probabilities for two outcomes j and k are limit-incommensurate if $m_j \neq m_k$. Furthermore, they are limit-incommensurate unless $\sum_{t\notin S_j} g(x_{it}) = \sum_{s\notin S_k} g(x_{is})$ (or, equivalently, unless $\sum_{t\in S_j} g(x_{it}) = \sum_{s\in S_k} g(x_{is})$). This is clearly impossible if $T = 3$ unless two of the $g(x_{it})$ are identical, just as when $T = 2$. Even for $T > 3$, we see that $\mathbf{p}_{x\theta}(\alpha)$ for the probit specification does not generally lie in a hyperplane except perhaps for certain discrete values of $\mathbf{x_i}$. Therefore the model is not identified unless the support of $\mathbf{x_i}$ contains those discrete values, and even in this case, if some element of $\mathbf{x_i}$ has a continuous distribution, then Theorem 2.3 applies and $\sqrt{n}$-consistent estimation is impossible.

3.4 Some identified models

This section catalogs some simple classes of models that are known to be identified, and shows how identification and estimation of each fits into the framework of this paper. They each have the strong property that $\mathbf{p}_{x\theta}(\alpha)$ lies in a hyperplane for all values of $\mathbf{x_i}$. This is important, because it allows consistent estimation that does not depend on $\mathbf{x_i}$ having support in some special region. In addition, in each of the following models identification is possible with only two time periods.

Additive separability (linear probability model)

One simple case in which identification can be achieved is when the fixed effect is additively separable within the probability function:

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = h(\alpha_i) + g(x_{it}, \theta).$$

The most popular example of this is the linear probability model, where $\Pr(y_{it} = 1) = \alpha_i + x_{it}\theta$. In this model the fixed effect can be removed using standard first differencing methods.

It is easy to verify that the probability function lies in a hyperplane, as long as the distribution of α_i is restricted such that $h(\alpha_i) + g(x_{it}, \theta)$ stays within [0,1]. If $T = 2$ there is a single hyperplane restriction is given by:

$$p_{10} - p_{01} = g(x_{i2}, \theta) - g(x_{i2}, \theta). \tag{3.6}$$

This can be rewritten as

$$E(y_1 - y_2|\mathbf{x_i}) = g(x_{i2}, \theta) - g(x_{i2}, \theta).$$

Which is the familiar condition underlying standard differencing methods.

Multiplicative separability (Poisson-quasi-CMLE)

Another known case is when the fixed effect is multiplicatively separable in the probability function:

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = h(\alpha_i)g(x_{it}, \theta).$$

Wooldridge (1999) shows how θ in this model can be consistently estimated using a quasi-CMLE, treating y_{it} as if it were a Poisson random variable with $E(y_{it}|\mathbf{x_i}, \alpha_i) = h(\alpha_i)g(x_{it}, \theta)$. We can verify that for this class of functional forms, $\mathbf{p}_{x\theta}(\alpha)$ once again lies in a hyperplane for all values of $\mathbf{x_i}$, assuming the $\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i)$ function does not go out of the [0,1] bounds. If $T = 2$ the hyperplane restriction can be written as:

$$\frac{p_{01} + p_{11}}{p_{10} + p_{11}} = \frac{g(x_{i2}, \theta)}{g(x_{i1}, \theta)}.$$

This condition can be rewritten as:

$$E\left(\frac{y_{i1}}{g(x_{i1})} - \frac{y_{i2}}{g(x_{i2})}|\mathbf{x_i}\right) = 0$$

Again this can be used to construct a method of moments estimator by using instruments that are functions of x_{it}.[8]

This model provides an example where $\sqrt{n}$-consistent estimation is possible, despite the fact that the no sufficient statistic for α_i exists (other than the data itself), so the conditioning approach is impossible.

Multiplicative separability in the odds ratio (logit)

A limitation of both models above is that they imply restrictions on either the joint support of x_{it} and α_i or the functions $h()$ and $g()$ in order to keep the predicted probabilities between 0 and 1, assumptions that may be unnatural in many applications.

A case that allows more natural assumptions is when α_i is multiplicatively separable in the odds ratio:

$$\frac{\Pr(y_{it}=1|\mathbf{x_i},\alpha_i)}{1-\Pr(y_{it}=1|\mathbf{x_i},\alpha_i)} = r(x_{it},\alpha_i,\theta) = h(\alpha_i)g(x_{it},\theta)$$

where the functions and/or distributions are restricted to keep $r()$ in the range $[0,\infty)$. We can write this model as:

$$\Pr(y_{it}=1|\mathbf{x_i},\alpha_i) = \frac{h(\alpha_i)g(x_{it},\theta)}{1+h(\alpha_i)g(x_{it},\theta)} \tag{3.7}$$

[8] Compare this formula to the first order conditions from the Poisson-Quasi-MLE:

$$\frac{1}{n}\sum\left(\frac{y_{i1}}{g(x_{i1})} - \frac{y_{i2}}{g(x_{i2})}\right)\left(\frac{g_\theta(x_{i1})g(x_{i2}) - g(x_{i1})g_\theta(x_{i2})}{g(x_{i1})+g(x_{i2})}\right) = 0.$$

where $g_\theta(x_{it})$ denotes $\frac{\partial g(x_{it})}{\partial\theta}$. Note that this Quasi-CMLE estimator does not have the usual efficiency properties, since the true distribution of y_{it} is by assumption not Poisson. The true optimal instruments will depend on the distribution of α_i. A two- step procedure could be used to improve the instruments and yield an estimator asymptotically more efficient than the Quasi-CMLE.

It is easy to see that this includes the standard logit specification (3.4). In fact, we can always re-parameterize equation (3.7) without loss of generality to look like a generalized form of the logit model with a possibly non-linear index:

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = \frac{e^{\alpha_i^* + g^*(x_{it},\theta)}}{1 + e^{\alpha_i^* + g^*(x_{it},\theta)}}$$

Once again this model obeys a hyperplane restriction for all values of $\mathbf{x_i}$. If $T = 2$, there is a single hyperplane restriction that can be written:

$$\frac{p_{01}}{g(x_{i2},\theta)} - \frac{p_{10}}{g(x_{i1},\theta)} = 0$$

We can rewrite this condition as follows:

$$E\left(\frac{(1-y_{i1})y_{i2}}{g(x_{i2})} - \frac{y_{i1}(1-y_{i2})}{g(x_{i1})}|\mathbf{x_i}\right) = 0$$

Once again this can be used to construct a method of moments estimator by using instruments that are functions of x_{it}.[9]

For the case of $T = 2$ the logit model has a unique property. Consider the class of models that meet the following condition, that I will call "unbounded types:"

$\lim_{\alpha_i \to \inf(A)} G(x, \theta_0, \alpha_i) = 0$

$\lim_{\alpha_i \to \sup(A)} G(x, \theta_0, \alpha_i) = 1.$

[9] In this case the optimal instruments do not depend on the distribution of α_i and can be calculated. Let $\psi(\theta) = \frac{(1-y_{i1})y_{i2}}{g(x_{i2})} - \frac{y_{i1}(1-y_{i2})}{g(x_{i1})}$. Then the optimal instruments are:

$$E(\partial\psi(\theta)/\partial\theta|\mathbf{x_i})[V(\psi(\theta)|\mathbf{x_i})]^{-1} = \left.\frac{g_\theta(x_{i1})g(x_{i2}) - g(x_{i1})g_\theta(x_{i2})}{g(x_{i1}) + g(x_{i2})}\right|_{\theta_0}$$

These are the same asymptotic weights implied by the first order conditions from the standard fixed-effects logit CMLE estimator, which can be written as:

$$\frac{1}{n}\sum\left(\frac{(1-y_{i1})y_{i2}}{g(x_{i2})} - \frac{y_{i1}(1-y_{i2})}{g(x_{i1})}\right)\left(\frac{g_\theta(x_{i1})g(x_{i2}) - g(x_{i1})g_\theta(x_{i2})}{g(x_{i1}) + g(x_{i2})}\right) = 0$$

They are also the same weights used by the Poisson-quasi-CMLE estimator.

This assumption says that the model allows the population to contain individuals for which the outcome is always 0 or always 1, even for extreme values of x_{it}. (It does not require the population to contain such types, only that the model does not rule them out.) A model that appears often in the literature that has this property is the linear index latent variable model (3.3), along with the assumption that α_i may take any value in $(-\infty, \infty)$.

For models in this class with $T = 2$, the logit assumption is the only one for which $\mathbf{p}_{x\theta}(\alpha)$ lies in a hyperplane for every possible value of x_{it}. The unbounded types assumption implies that $\mathbf{p}_{x\theta}(\alpha)$ comes arbitrarily close to the origin (where the probability of all zeros is 1), and to the point where $p_{11} = 1$. By Lemma 2.4, the necessary condition is that there must exist constants $\{c_{01}, c_{10}, c_{11}\} \neq \mathbf{0}$ such that for all values of $\alpha_i \in A$,

$$c_{01}p_{01} + c_{10}p_{10} + c_{11}p_{11} = 0 \tag{3.8}$$

But since the unbounded types assumption ensures that $\mathbf{p}_{x\theta}(\alpha)$ comes arbitrarily close to the point where $p_{11} = 1$, we must have $c_{11} = 0$. Therefore equation (3.8) requires $p_{01}/p_{10} = -c_{10}/c_{01}$, i.e. the ratio is a function of $\mathbf{x_i}$ only, and does not depend on α_i. Because of the independence assumption,

$$\frac{p_{01}}{p_{10}} = \frac{(1 - G(x_{i1}, \alpha_i, \theta))G(x_{i2}, \alpha_i, \theta)}{G(x_{i1}, \alpha_i, \theta)(1 - G(x_{i2}, \alpha_i, \theta))} = \frac{r(x_{i1}, \alpha_i, \theta)}{r(x_{i2}, \alpha_i, \theta)}$$

where $r(\cdot)$ represents the odds ratio $G(\cdot)/(1 - G(\cdot))$.

So the "unbounded types" assumption requires that

$$\frac{r(x_{i1}, \alpha_i, \theta)}{r(x_{i2}, \alpha_i, \theta)}$$

doesn't depend on α_i. For this to be true for all values of $\mathbf{x_i}$, $r()$ must be multiplicatively separable, implying the generalized logit form. [10]

[10] Chamberlain (1992) apparently proved something very similar to this, but I have not yet been able to obtain a copy of his unpublished paper.

Comments

In each of the specifications above, the necessary condition for identification places little restriction on the functional form of the $g()$ function. In fact, given sufficient variation in $x_{i2} - x_{i1}$, we can often nonparametrically identify the $g()$ function up to a level or scale parameter. The identification results presented here depend on how the individual effects α_i enter, placing little restriction on the form of how $\mathbf{p}(\mathbf{x_i}, \alpha_i, \theta)$ varies with $\mathbf{x_i}$.

3.5 Models with additional parameters

Time specific coefficients on individual effects (factor loadings)

Much like we did for the probit model, we can use the approach in Lemma 2.5 to check for identification in variations of the binary panel data model that have additional parameters. Consider a binary choice latent variable model with time specific coefficients on the fixed effects:

$$y_{it}^* = \delta_t \alpha_i + g(x_{it}, \theta) + \varepsilon_t, \; y_{it} = \mathbf{1}(y_{it}^* > 0)$$

where δ_t is a positive parameter that may take a different value for each time period, and α_i may take any value in $(-\infty, \infty)$. Arellano and Honore (2001) briefly discus a model with such coefficients, which they call "factor loadings." They observe that neither the CMLE nor the maximum score approaches will work, but they state that it is "less clear" if such a model is identified. Using the approach in this paper, it is simple to show that the answer is negative.

Consider two possible outcomes j and k. Then,

$$\lim_{\alpha_i \to -\infty} \frac{p_j}{p_k} = \lim_{\alpha_i \to -\infty} \frac{\prod_{t \in S_j} G(\delta_t \alpha_i + g(x_{it}))}{\prod_{s \in S_k} G(\delta_s \alpha_i + g(x_{is}))}$$

where $G()$ is the CDF of ε_t. Because $\lim_{\alpha_i \to -\infty} G(\delta_t \alpha_i + g(x_i)) = 0$, Lemma 2.4 applies, so we need to check whether $\mathbf{p}_{x\theta}(\alpha)$ is linearly dependent.

It will suffice to examine the case where ε_t is logistic. In this case

$$\Pr(y_{it} = 1|\mathbf{x_i}, \alpha_i) = \frac{e^{\delta_t \alpha_i + g(x_{it},\theta)}}{1 + e^{\delta_t \alpha_i + g(x_{it},\theta)}}$$

so we have

$$\lim_{\alpha_i \to -\infty} \frac{p_j}{p_k} = \lim_{\alpha_i \to -\infty} \frac{\prod_{t \in S_j} e^{\delta_t \alpha_i + g(x_{it})}}{\prod_{s \in S_k} e^{\delta_s \alpha_i + g(x_{is})}} = \lim_{\alpha_i \to -\infty} \frac{e^{\alpha_i \sum_{t \in S_j} \delta_t + \sum_{t \in S_j} g(x_{it})}}{e^{\alpha_i \sum_{s \in S_k} \delta_s + \sum_{s \in S_k} g(x_{is})}}.$$

Clearly the two probabilities will be limit-incommensurate unless $\sum_{t \in S_j} \delta_t = \sum_{s \in S_k} \delta_s$. Since it normally wouldn't make sense to impose this restriction, $\mathbf{p}_{x\theta}(\alpha)$ is not linearly dependent by Lemma 2.5, and θ is not identified. A similar result will hold if we choose the distribution of ε_{it} to be any distribution that has a density of the form

$$f(x) = ax^b exp(-cx^d),\ d > 0,$$

a class that in addition to the logistic includes the normal, gamma (including exponential and chi-squared), and weibull distributions. For this class, it can be shown that:

$$\lim_{x \to \infty} \frac{x^{b-d+1} exp(-cx^d)/cd}{1 - F(x)} = 1$$

and so in the limit the ratio of probabilities will still behave like an exponential function, and the same non-identification result will follow in a similar way. Of course in the absence of any distributional assumption, θ remains non-identified.

Random trend model

Consider a latent variable model in which in addition to an individual fixed effect α_i in the index function, there is an unobserved individual time trend β_i:

$$y^*_{it} = \alpha_i + \beta_i t + g(x_{it}, \theta) + \varepsilon_t,\ y_{it} = \mathbf{1}(y^*_{it} > 0).$$

where both α_i may and β_i may take any value in $(-\infty, \infty)$. If y^*_{it} were observed, and we had at least 3 time periods, we could estimate θ by twice

differencing to remove both individual effects. If only the discrete outcome y_{it} is observed, at least 4 periods are required for identification, as I will now show. Again, it suffices to consider the case where ε_t has a logistic distribution.

Because there are now two unobserved effects for each individual, this model doesn't precisely fit the assumptions of Theorem 2.2. We can still use the theorem by assuming that one of the two fixed effects is observed, and considering the functions $\mathbf{p}_{x\theta\beta}(\alpha_i)$ or $\mathbf{p}_{x\theta\alpha}(\beta_i)$. If there is no restriction on the joint distribution of α_i and β_i, then a necessary condition for identification is that there must be some value of $\mathbf{x_i}$ such that both these functions must lie in a hyperplane, for all values of α_i and β_i.

Consider two different outcomes j and k:

$$\begin{aligned}
\frac{p_j}{p_k} &= \frac{\prod_{t\in S_j} e^{\alpha_i+\beta_i t+g(x_{it})}}{\prod_{s\in S_k} e^{\alpha_i+\beta_i t+g(x_{is})}} \\
&= \frac{e^{\alpha_i m_j+\beta_i \sum_{t\in S_j} t+\sum_{t\in S_j} g(x_{it})}}{e^{\alpha_i m_k+\beta_i \sum_{s\in S_k} s+\sum_{s\in S_k} g(x_{is})}} \\
&= e^{\alpha_i(m_j-m_k)+\beta_i(\sum_{t\in S_j} t-\sum_{s\in S_k} s)+\sum_{t\in S_j} g(x_{it})-\sum_{s\in S_k} g(x_{is})}
\end{aligned}$$

Thus the two outcomes will be limit-incommensurate unless they have the same number of "1" outcomes ($m_j = m_k$), and unless these "1" outcomes come in the periods such that $\sum_{t\in S_j} t = \sum_{s\in S_k} s$. Clearly this is impossible if T=3, so unlike the continuous model, identification is impossible when T=3. However, if T=4, then the outcomes (1001) and (0110) will not be limit-incommensurate. In fact we can write:

$$\frac{p_{1001}}{p_{0110}} = \exp(g(x_{i1},\theta) - g(x_{i2},\theta) - g(x_3,\theta) + g(x_4,\theta))$$

which does not depend on α_i. Thus, given sufficient variation in the x variables, θ can be estimated by weighted least squares imposing this moment restriction, or by using conditional maximum likelihood on the subset of observations for which the outcome is (1001) or (0110).

Heteroskedasticity

Consider a latent variable model in which in addition to an individual fixed effect α_i in the index function, the disturbance term has individual specific variance:

$$y_{it}^* = \alpha_i + g(x_{it}, \theta) + \sigma_i \varepsilon_t,\ y_{it} = \mathbf{1}(y_{it}^* > 0).$$

where σ_i may take any value in $(0, \infty)$. Again, it suffices to consider the case where ε_t has a logistic distribution.

Consider two different outcomes j and k:

$$\begin{aligned} \frac{p_j}{p_k} &= \frac{\prod_{t \in S_j} e^{(\alpha_i + g(x_{it}))/\sigma_i}}{\prod_{s \in S_k} e^{(\alpha_i + g(x_{is}))/\sigma_i}} \\ &= e^{\frac{1}{\sigma_i}(\alpha_i(m_j - m_k) + \sum_{t \in S_j} g(x_{it}) - \sum_{s \in S_k} g(x_{is}))} \end{aligned}$$

Thus the two outcomes will be limit-incommensurate unless they have the same number of "1" outcomes ($m_j = m_k$), and unless $\sum_{t \in S_j} g(x_{it}) = \sum_{s \in S_k} g(x_{is}$. This is much like the fixed effects probit case: identification is only possible using the maximum-score principle, and $\sqrt{n}$-consistent estimation is impossible.

3.6 Another identifiable functional form when T $\geq$ 3

Section 3.4 presented three classes of functional forms that obey a hyperplane restriction for every value of $\mathbf{x_i}$, and that allow identification when $T = 2$. However, for T$\geq$3, these forms impose much stronger restrictions than are required for identification by Theorem 2.2. In fact they impose several independent hyperplane restrictions, when only one is required. The question remains whether there are more flexible functional forms besides those in section 3.4 that exhibit the property of having $\mathbf{p}_{x\theta}(\alpha)$ lie in a hyperplane for every value of $\mathbf{x_i}$.

Consider only the class of specifications where α_i is additively separable within the latent variable model, so that $\Pr(y_{it} = 1) = G(\alpha_i + g(x_{it}, \theta))$. This is not as restrictive as it might first appear, because we can re-parameterize

the model in various ways. If we assume that $A = (-\infty, \infty)$, then the "unbounded types" assumption holds and identification requires that $\mathbf{p}_{x\theta}(\alpha)$ be linearly dependent.

We will look for functional forms that obey a hyperplane restriction of one the following forms:

$$c_{001}p_{001} + c_{010}p_{010} + c_{100}p_{100} = 0 \tag{3.9}$$

$$c_{011}p_{011} + c_{110}p_{110} + c_{101}p_{101} = 0. \tag{3.10}$$

Assume the first case is true, divide through by p_{000}, and rearrange to get

$$k_1 r(\alpha_i + g(x_{i1}, \theta)) + k_2 r(\alpha_i + g(x_{i2}, \theta)) = r(\alpha_i + g(x_3, \theta)) \tag{3.11}$$

where k_1, k_2 are constants, and $r(\cdot)$ is the odds ratio $G(\cdot)/(1 - G(\cdot))$.

Now consider the choice of $\mathbf{x_i}$ such that $g(x_{i2}, \theta) - g(x_{i1}, \theta) = g(x_3, \theta) - g(x_{i2}, \theta) = \Delta$. In other words, the $g()$ index functions are evenly spaced. If $\mathbf{p}_{x\theta}(\alpha)$ is linearly dependent for all $\mathbf{x_i}$, it must also be linearly dependent for this specific choice of $\mathbf{x_i}$. Rewrite equation (3.11) as

$$k_1 r(\alpha_i + g(x_{i1}, \theta)) + k_2 r(\alpha_i + g(x_{i1}, \theta) + \Delta) = r(\alpha_i + g(x_{i1}, \theta) + 2\Delta).$$

Since this equation must hold for any choice of α_i, it must be true when $\alpha_i = n\Delta - g(x_{i1}, \theta)$ for any n. So it must be that:

$$k_1 r(n\Delta) + k_2 r((n+1)\Delta) = r((n+2)\Delta)$$

Write this expression as a difference equation, $s_{n+2} = k_1 s_n + k_2 s_{n+1}$, where $s_n = r(n\Delta)$. The solutions of this difference equation have the general form $s_n = a_1 b_1^n + a_2 b_2^n$. Therefore

$$r(x, \alpha_i, \theta) = a_1 b_1^{\alpha_i + g(x,\theta)} + a_2 b_2^{\alpha_i + g(x,\theta)}$$

which we can re-parameterize without loss of generality to get:

$$r(x, \alpha_i, \theta) = \lambda_1^{\alpha_i + g(x,\theta)} + c\lambda_2^{\alpha_i + g(x,\theta)}$$

Solving for the probability function $G()$ gives us:

$$G(\alpha_i + g(x,\theta)) = \frac{\lambda_1^{\alpha_i+g(x,\theta)} + c\lambda_2^{\alpha_i+g(x,\theta)}}{1 + \lambda_1^{\alpha_i+g(x,\theta)} + c\lambda_2^{\alpha_i+g(x,\theta)}}$$

Logit is a special case of this, with $\lambda_1 = e$ and $c = 0$.

If we assume instead that $c_{011}p_{011} + c_{110}p_{110} + c_{101}p_{101} = 0$, divide through by p_{111}, and follow similar reasoning, we get:

$$\frac{1}{r(x,\alpha_i,\theta)} = \lambda_1^{\alpha_i+g(x,\theta)} + c\lambda_2^{\alpha_i+g(x,\theta)}$$

$$G(\alpha_i + g(x,\theta)) = \frac{1}{1 + \lambda_1^{\alpha_i+g(x,\theta)} + c\lambda_2^{\alpha_i+g(x,\theta)}}$$

which is really the same functional form as before, but with opposite conventions for labelling the outcomes 1 or 0.

We derived this functional form to be linearly dependent for choices of $\mathbf{x_i}$ such that $g(x_{i2},\theta) - g(x_{i1},\theta) = g(x_3,\theta) - g(x_{i2},\theta)$. It can easily be shown that $\mathbf{p}_{x\theta}(\alpha)$ for this form is linearly dependent for *all* possible choices of $\mathbf{x_i}$. If we start with the assumption in equation (3.9), we can write the hyperplane restriction as:

$$(\lambda_2^{\Delta_2} - \lambda_1^{\Delta_2})p_{100} + (\lambda_1^{\Delta_2}\lambda_2^{-\Delta_1} - \lambda_1^{-\Delta_1}\lambda_2^{\Delta_2})p_{010} + (\lambda_1^{-\Delta_1} - \lambda_2^{-\Delta_1})p_{001} = 0$$

where $\Delta_1 = g(x_{i2},\theta) - g(x_{i1},\theta)$, and $\Delta_2 = g(x_{i3},\theta) - g(x_{i2},\theta)$.

As in previous cases, estimates could often be obtained by imposing this moment condition in the sample. Notice that this condition does not depend on the constant c. I assume that λ_1 and λ_2 are known, but it may also be possible to estimate these parameters, although it would likely be difficult numerically. Clearly the constant c cannot be estimated this way. These functional forms are quite similar to logit for most purposes, and are probably of little use to practitioners.

4 Concluding Remarks

The main contribution of this paper is to present a general approach to checking for identification in discrete choice panel data models. As seen in Section 3.4, the approach brings together previous results and provides a unifying framework, and some intuition about where the identification is coming from. Sections 3.5 and 3.6 show how the approach can be used both to prove non-identification, and to suggest new models that may be identified.

For the examples in Section 3, we were able to rule out identification using limit-incommensurability as in Lemma 2.5. In other cases this approach may not work. Linear dependence of the function $\mathbf{p}_{x\theta}(\alpha)$ requires that the determinant:

$$\begin{vmatrix} p_1(\mathbf{x},\theta,\alpha_1) & ... & p_{K-1}(\mathbf{x},\theta,\alpha_1) \\ \vdots & & \vdots \\ p_1(\mathbf{x},\theta,\alpha_{K-1}) & ... & p_{K-1}(\mathbf{x},\theta,\alpha_{K-1}) \end{vmatrix}$$

is equal to zero for all possible values of $\{\alpha_1, ..., \alpha_{K-1}\}$. For some functional forms it may be possible to show non-identification by computing this determinant numerically. For others, such as probit, that are "close" to identified, it may be hard to distinguish a true zero determinant from numerical error.

This paper deals only with strict identification of the common parameters in a parametric model. As suggested in section 3, it may be possible to place tight bounds on parameters even when a model is not identified. (See Honore and Tamer (2003) for a related discussion of bounds in random effects models).

This paper also does not deal directly, with estimation of average marginal effects across the distribution of α_i in the population. (This is a weakness of fixed-effects models in general.) Although assuming a logit specification is enough to identify the parameters in a panel logit model, it is clear that it is not enough to identify the marginal effects. On the other hand, it is clear that it would be possible to place some bounds on the sizes of the marginal effects. For example, because the probability derivative for an observation in a logit model is given by $\beta P(1-P)$, a very simple upper bound on the

population average probability derivative for a linear model is 0.25β. The approach in this paper suggests ways that tighter bounds than these could be identified in principle (including lower bounds). More examination of this issue is a goal of future research.

A Appendix

Proof of Theorem 2.3

Following Chamberlain (1986), we first specify a class of parametric models $\lambda(\delta)$ for the distribution of α_i, where δ is a unidimensional parameter. Let the true distribution $F(\alpha|\mathbf{x})$ be a discrete distribution with a finite number of points α_m^o, each taken with probability $w_m^o > 0$ (similar to the proof to Theorem 2.2), where both α_m^o and w_m^o may depend on $\mathbf{x}$. Then the true reduced form $\mathbf{p_o^*}(\mathbf{x}) = \sum_{m=1}^K w_m^o \mathbf{p}(\mathbf{x}, \theta_0, \alpha_m^o)$. Assume that for each $\mathbf{x} \in \mathcal{X}_n$, $\mathbf{p_o^*}(\mathbf{x})$ is at the center of some open ball of radius $\epsilon > 0$ entirely within the convex hull $H_{x\theta_0}$. Note that this also implies that all elements of $\mathbf{p_o^*}(\mathbf{x})$ are bounded away from zero.

For any $\mathbf{x} \in \mathcal{X}_n$, we can always choose a finite set of values α_m^o such that the $K-1$ dimensional space is spanned by affine combinations of $\mathbf{p}(\mathbf{x}, \theta_0, \alpha_m^o)$. Therefore for any element θ_j of θ, we can choose affine weights w_m' (which may depend on $\mathbf{x}$) such that:

$$\sum_{m=1}^K w_m' \mathbf{p}(\mathbf{x}, \theta_0, \alpha_m^o) = \sum_{m=1}^K w_m^o [\mathbf{p}(\mathbf{x}, \theta_0, \alpha_m^o) - \partial \mathbf{p}(\mathbf{x}, \theta_0, \alpha_m^o)/\partial \theta_j]$$

Since the derivatives are continuous and $\mathcal{X}_n$ is compact, the weights w_m' are bounded. For $\mathbf{x} \in \mathcal{X}_n$ the parametric submodel is given by:

$$\mathbf{p}^*(\mathbf{x}, \theta, \delta) = \sum_{m=1}^K (w_m^o(1-\delta) + \delta w_m') \mathbf{p}(\mathbf{x}, \theta, \alpha_m^o).$$

The weights $(w_m^o(1-\delta) + \delta w_m')$ will be strictly positive and will sum to 1 in the neighborhood of the true value $\delta_0 = 0$, so they specify a valid discrete distribution for α. The weights are constructed so that at (θ_0, δ_0) a small change in δ has the same effect as a small change in θ_j. For $\mathbf{x} \notin \mathcal{X}_n$ we can simply let $F(\alpha|\mathbf{x})$ be a degenerate distribution at α', so $\mathbf{p}^*(\mathbf{x}) = \mathbf{p}(\mathbf{x}, \theta, \alpha')$ for some value α'.

The conditional likelihood function for a single outcome d is given by:

$$f(d|\mathbf{x},\theta,\delta) = \sum_{k=1}^{K} \mathbf{1}(d=\omega_k) p_k^*(\mathbf{x},\theta,\delta)$$

Next we show that $f(d|\mathbf{x},\theta,\delta)$ is mean-square differentiable. It suffices to show that $p_k^*(\mathbf{x},\theta,\delta)$ is mean-square differentiable. In other words,

$$p_k^{*1/2}(\mathbf{x},\theta+h,\delta) - p_k^{*1/2}(\mathbf{x},\theta,0) = h'\psi_\theta + \delta\psi_\delta + r(\mathbf{x},h,\delta)$$

such that

$$\lim_{h\to 0,\delta\to 0} \int r(\mathbf{x},h,\delta)^2 d\mu/(\|h\|+|\delta|)^2 = 0 \tag{A.1}$$

with

$$\psi_\theta = \frac{1}{2} p_k^{*-1/2}(\mathbf{x},\theta_0,0)\partial p_k^*(\mathbf{x},\theta_0,0)/\partial\theta$$

$$\psi_\delta = \frac{1}{2} p_k^{*-1/2}(\mathbf{x},\theta_0,0)\partial p_k^*(\mathbf{x},\theta_0,0)/\partial\delta$$

A sufficient condition for mean square differentiability is that there exists functions $q_j(\mathbf{x})$ and $q_\delta(\mathbf{x})$ such that $|\psi_{\theta_j}| < q_j(\mathbf{x})$ and $|\psi_\delta| < q_\delta(\mathbf{x})$ in a neighborhood of θ_0, δ_0, with $E(q_j^2(\mathbf{x})) < \infty$ and $E(q_\delta^2(\mathbf{x})) < \infty$. Then condition (A.1) holds by the dominated convergence theorem.

For $\mathbf{x} \in \mathcal{X}_n$ we have:

$$\psi_\theta = \frac{1}{2} p_k^{*-1/2}(\mathbf{x},\theta_0,0) \sum_{m=1}^{K} w_m^o \partial p_k(\mathbf{x},\theta_0,\alpha_m^o)/\partial\theta$$

$$\psi_\delta = \frac{1}{2} p_k^{*-1/2}(\mathbf{x},\theta_0,0) \sum_{m=1}^{K} (w_m' - w_m^o) p_k(\mathbf{x},\theta_0,\alpha_m^o)$$

These are bounded, since $p_k^*(\mathbf{x},\theta_0,0)$ is bounded away from zero by assumption, $\partial p_k(\mathbf{x},\theta_0,\alpha_m^o)/\partial\theta$ is continuous, and $\mathcal{X}_n$ is compact.

For $\mathbf{x} \notin \mathcal{X}_n$ we have:

$$\psi_\theta = \frac{1}{2} p_k^{-1/2}(\mathbf{x},\theta_0,\alpha')\partial p_k(\mathbf{x},\theta_0,\alpha')/\partial\theta,\ \psi_\delta = 0$$

which is bounded by the assumption in the theorem.

To show that the semiparametric information bound for some element θ_j of θ is zero, it now suffices to show that we can choose a submodel to make $E(\psi_{\theta_j} - \psi_\delta)^2$ arbitrarily close to zero. This is the case, since

$$E(\psi_{\theta_j} - \psi_\delta)^2 = E((\psi_{\theta_j} - \psi_\delta)^2 | \mathbf{x} \in \mathcal{X}_n) \Pr(\mathbf{x} \in \mathcal{X}_n)$$

$$+ \; E(\psi_{\theta_j}^2 | \mathbf{x} \notin \mathcal{X}_n) \Pr(\mathbf{x} \notin \mathcal{X}_n)$$

For $\mathbf{x} \in \mathcal{X}_n$, the difference is zero by construction, and for $\mathbf{x} \notin \mathcal{X}_n$, ψ_{θ_j} is bounded by assumption, and the probability $\Pr(\mathbf{x} \notin \mathcal{X}_n)$ can be made arbitrarily small.

□

References

Altonji, J. G., and R. L. Matzkin (2001): "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *NBER Technical Working Paper*, 267.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, 32, 283–301.

Arellano, M., and B. E. Honore (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 5, pp. 3229–3296. North Holland, Amsterdam.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics, Vol. 2*, ed. by Z. Griliches, and M. Intriligator, chap. 22, pp. 1247–1318. Amsterdam: North Holland.

——— (1986): "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

——— (1992): "Binary Response Models for Panel Data: Identification and Information," Harvard Univ., unpublished manuscript.

Honore, B. E. (2002): "Nonlinear Models with Panel Data," Cemmap Working Paper CWP13/02.

Honore, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70,(5), 2053–63.

Honore, B. E., and E. Tamer (2003): "Bounds on Parameters in Dynamic Discrete Choice Models," working paper.

Johnson, E. G. (2004): "Panel Data Models with Discrete Dependent Variables," Ph.D. thesis, Stanford Univeristy.

Lay, S. R. (1982): *Convex Sets and their Applications*. John Wiley and Sons, Inc.

Magnac, T. (Forthcoming): "Binary Variables Models and Sufficiency: Generalizing Conditional Logit," *Econometrica*.

Manski, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), 357–62.

——— (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83(403), 729–738.

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

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

Wooldridge, J. M. (1999): "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics*, 90, 77–97.

——— (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.