What's New in Econometrics? NBER, Summer 2007
Lecture 4, Monday, July 30th, 3.15-4.15 pm

Nonlinear Panel Data Models

These notes summarize some recent, and perhaps not-so-recent, advances in the estimation of nonlinear panel data models. Research in the last 10 to 15 years has branched off in two directions. In one, the focus has been on parameter estimation, possibly only up to a common scale factor, in semiparametric models with unobserved effects (that can be arbitrarily correlated with covariates.) Another branch has focused on estimating partial effects when restrictions are made on the distribution of heterogeneity conditional on the history of the covariates. These notes attempt to lay out the pros and cons of each approach, paying particular attention to the tradeoff in assumptions and the quantities that can be estimated.

1. Basic Issues and Quantities of Interest

Most microeconomic panel data sets are best characterized as having few time periods and (relatively) many cross section observations. Therefore, most of the discussion in these notes assumes T is fixed in the asymptotic analysis while N is increasing. We assume random sample in the cross section, $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$. Take y_{it} to be a scalar for simplicity. If we are not concerned about traditional (contemporaneous) endogeneity, then we are typically interested in

$$D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \tag{1.1}$$

or some feature of this distribution, such as $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or a conditional median. In the case of a mean, how do we summarize the partial effects? Let $m_t(\mathbf{x}_t, \mathbf{c})$ be the mean function. If x_{tj} is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}, \tag{1.2}$$

or look at discrete changes. How do we account for unobserved $\mathbf{c}_i$? If we want to estimate magnitudes of effects, we need to know enough about the distribution of $\mathbf{c}_i$ so that we can insert meaningful values for $\mathbf{c}$. For example, if $\boldsymbol{\mu}_{\mathbf{c}} = E(\mathbf{c}_i)$, then we can compute the *partial effect at the average (PEA)*,

$$\theta_j(\mathbf{x}_t, \boldsymbol{\mu}_{\mathbf{c}}). \tag{1.3}$$

Of course, we need to estimate the function m_t and the mean of $\mathbf{c}_i$. If we know more about the distribution of $\mathbf{c}_i$, we can insert different quantiles, for example, or a certain number of standard deviations from the mean.

Alternatively, we can average the partial effects across the distribution of $\mathbf{c}_i$:

$$\text{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)]. \tag{1.4}$$

The difference between (1.3) and (1.4) can be nontrivial for nonlinear mean functions. The definition in (1.4) dates back at least to Chamberlain (1982), and is closely related to the notion of the average structural function (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$\text{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \tag{1.5}$$

Assuming the derivative passes through the expectation results in (1.5), the average partial effect. Of course, computing discrete changes gives the same result always. APEs are directly across models, and APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Semiparametric methods, which, by construction, are silent about the distribution of $\mathbf{c}_i$, unconditionally or conditional on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, cannot generally deliver estimates of average partial (marginal) effects. Instead, an index structure is usually imposed so that parameters can be consistently estimated. So, for example, with scalar heterogeneity we might have an index model with additive heterogeneity:

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.6}$$

where, say, $G(\bullet)$ is strictly increasing and continuously differentiable (and, in some cases, is known, and in others, is not). Then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.7}$$

where $g(\bullet)$ is the derivative of $G(\bullet)$. Then estimating β_j means we can estimate the sign of the partial effect, and even the relative effects of any two continuous variables. But, even if $G(\bullet)$ is specified (the more common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i: the size of the scale factor for a random draw i, $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$, depends on c_i. Without knowing something about the distribution of c_i we cannot generally locate $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$.

Returning to the general case, Altonji and Matzkin (2005) focus on what they call the *local average response (LAR)* as opposed to the APE or PAE. The LAR at $\mathbf{x}_t$ for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t), \tag{1.8}$$

where $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a "local" partial effect because it averages out the heterogeneity for the slice of the population given by the vector $\mathbf{x}_t$. The APE, which by comparison could be called a "global average response," averages out over the entire distribution of $\mathbf{c}_i$. See also Florens, Heckman, Meghir, and Vytlacil (2004).

It is important to see that the definitions of partial effects does not depend on the nature of the variables in $\mathbf{x}_t$ (except for whether it makes sense to use the calculus approximation or use changes). In particular, $\mathbf{x}_t$ can include lagged dependent variables and lags of other variables, which may or may not be strictly exogenous. Unfortunately, we cannot identify the APEs, or even relative effects in index models, without some assumptions.

2. Exogeneity Assumptions on the Covariates

Ideally, we would only have to specify a model for $D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$ or some feature. Unfortunately, it is well known that specifying a full parametric model is not sufficient for identifying either the parameters of the model or the partial effects defined in Section 1. In this section, we consider two useful exogeneity assumptions on the covariates.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i), \tag{2.1}$$

which means that $\mathbf{x}_{ir}$, $r \neq t$, does not appear in the conditional distribution of $\mathbf{y}_{it}$ once $\mathbf{x}_{it}$ and $\mathbf{c}_i$ have been counted for. Chamberlain (1984) labeled (2.1) *strict exogeneity conditional on the unobserved (or latent) effects* $\mathbf{c}_i$. Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i), \tag{2.2}$$

which we already saw for linear models. (In other cases a condition stated in terms of conditional medians is more convenient.) Of course, either version of the assumption rules out lagged dependent variables, but also other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in $\mathbf{x}_{ir}$, $r > t$. But it is the assumption underlying the most common estimation methods for nonlinear models.

More natural is a *sequential exogeneity* assumption (conditional on the unobserved effects) assumption, which we can state generally as

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{it},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i) \tag{2.3}$$

or, sometimes, in terms of specific features of the distribution. Assumption (2.3) allows for lagged dependent variables and does not restrict feedback. Unfortunately, it is much more difficult to allow, especially in nonlinear models.

Neither (2.2) nor (2.3) allows for contemporaneous endogeneity of one or more elements of

$\mathbf{x}_{it}$, where, say, x_{itj} is correlated with unobserved, time-varying unobservables that affect y_{it}, or where x_{itj} is simultaneously determined along with y_{it}. This will be covered in later notes on control function methods.

3. Conditional Independence Assumption

In some cases – certainly traditionally – a conditional independence assumption is imposed. We can write the condition generally as

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i,\mathbf{c}_i). \tag{3.1}$$

This assumption is only useful in the context of the strict exogeneity assumption (2.1), in which case we can write

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \tag{3.2}$$

In a parametric context, the conditional independence assumption therefore reduces our task to specifying a model for $D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$. In random effects and CRE frameworks, conditional independence plays a critical role in being able to estimate the parameters in distribution the of $\mathbf{c}_i$. We could get by with less restrictive assumptions by parameterizing the dependence, but that increases computational burden. As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. (That is, we do not need to place restrictions on $D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i)$.) Before we can study estimation, we must discuss the critical issue of the dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$.

4. Assumptions about the Unobserved Heterogeneity

The modern approach to panel data analysis with micro data treats the unobserved heterogeneity as random draws along with the observed data, and that is the view taken here. Nevertheless, there are still reasons one might treat them as parameters to estimate, and we allow for that in our discussion.

Random Effects

For general nonlinear models, what we call the "random effects" assumption is independence between $\mathbf{c}_i$ and $\mathbf{x}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$:

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{4.1}$$

If we combine this assumption with a model for $m_t(\mathbf{x}_t, \mathbf{c})$, then the APEs are actually nonparametrically identified. (And, in fact, we do not need to assume strict or sequential exogeneity to use a pooled estimation method, or to use just a single time period.) In fact, if $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ and $D(\mathbf{c}_i|\mathbf{x}_{it}) = D(\mathbf{c}_i)$, then the APEs are obtained from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{4.2}$$

(The argument is a simple application of the law of interated expectations; it is discussed in detail in Wooldridge (2005a).) In principle, $E(y_{it}|\mathbf{x}_{it})$ can be estimated nonparametrically, and we only need a single time period to identify the partial effects for that time period.

In some leading cases (for example random effects probit and Tobit with heterogeneity normally distributed), if we want to obtain partial effects for different values of $\mathbf{c}$, we must assume more: the strict exogeneity assumption (2.1), the conditional independence assumption (3.1), and the random effects assumption (4.1) with a parametric distribution for $D(\mathbf{c}_i)$ are typically sufficient. We postpone this discussion because it takes us into the realm of specifying parametric models.

Correlated Random Effects

A "correlated random effects" framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but the dependence in restricted in some way. In a parametric setting, we specify a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as in Mundlak (1978), Chamberlain (1982), and many subsequent authors. For many models, including probit and Tobit, one can allow $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend in a "nonexchangeable" manner on the time series of the covariates; Chamberlain's random effects probit model does this. But the distributional assumptions that lead to simple estimation – namely, homoskedastic normal with a linear conditional mean — are restrictive. But it is aslo possible to assume

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i) \tag{4.3}$$

without specifying $\mathrm{D}(c_i|\bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. We will see in the next section that (4.3) is very powerful.

We can go further. For example, suppose that we think the heterogeneity $\mathbf{c}_i$ is correlated with features of the covariates other than just the time average. Altonji and Matzkin (2005) allow for $\bar{\mathbf{x}}_i$ in equation (4.3) to be replaced by other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as sample variances and covariance. These are examples of "exchangeable" functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ – that is, statistics whose value is the same regardless of the ordering of the

$\mathbf{x}_{it}$. Non-exchangeable functions can be used, too. For example, we might think that $\mathbf{c}_i$ is correlated with individual-specific trends, and so we obtain $\mathbf{w}_i$ as the intercept and slope from the unit-specific regressions $\mathbf{x}_{it}$ on $1,\ t, t = 1,\ldots,T$ (for $T \geq 3$); we can also add the error variance from this individual specific regression if we have sufficient time periods. Then, the condition becomes

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\mathbf{w}_i). \tag{4.4}$$

Practically, we need to specify $\mathbf{w}_i$ and then establish that there is enough variation in $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ separate from $\mathbf{w}_i$; this will be clear in the next section.

Fixed Effects

Unfortunately, the label "fixed effects" is used in different ways by different researchers (and, sometimes, by the same researcher). The traditional view was that a fixed effects framework meant $\mathbf{c}_i,\ i = 1,\ldots,N$ were treated as parameters to estimate. This view is still around, and, when researchers say they estimated a nonlinear panel data model by "fixed effects," they sometimes mean the $\mathbf{c}_i$ were treated as parameters to estimate along with other parameters (whose dimension does not change with N). As is well known, except in special cases, estimation of the $\mathbf{c}_i$ generally introduces an "incidental parameters" problem. (More on this later when we discuss estimation methods, and partial effects.) Subject to computational feasibility, the approach that treats the $\mathbf{c}_i$ as parameters is widely applicable.

The "fixed effects" label can mean that $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted. Even in that case, there are different approaches to estimation of parameters. One is to specify a joint distribution $D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i)$ such that a sufficient statistic, say $\mathbf{s}_i$, can be found such that

$$D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{s}_i), \tag{4.5}$$

and where the latter distribution still depends on the parameters of interest in a way that identifies them. In most cases, the conditional independence assumption (3.1) is maintained, although one CMLE is known to have robustness properties: the so-called "fixed effects" Poisson estimator. We cover that later on.

5. Nonparametric Identification of Average Partial and Local Average Effects

Before considering identification and estimation of parameters in parametric models, it is useful to ask which quantities, if any, are identified without imposing parametric assumptions. Not surprisingly, there are no known results on nonparametric identificiation of partial effects evaluated at specific values of **c**, such as $\boldsymbol{\mu}_{\mathbf{c}}$ – except, of course, when the partial effects do not depend on **c**. Interestingly, identification can fail even under a full set of strong parametric assumptions. For example, in the probit model

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\boldsymbol{\beta} + c), \tag{5.1}$$

where **x** is $1 \times K$ an includes unity, the partial effect for a continuous variable x_j is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta} + c)$. The partial effect at the mean of c is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$. Suppose we assume that $c|\mathbf{x}$ ~Normal$(0, \sigma_c^2)$. Then it is easy to show that

$$P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}), \tag{5.2}$$

which means that only the scaled parameter vector $\boldsymbol{\beta}_c \equiv \boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}$ is identified. Therefore, $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$, is evidently unidentified. (The fact that probit of y on **x** estimates $\boldsymbol{\beta}_c$ has been called the "attenuation bias" that results from omitted variables in the context of probit, even when the omitted variable is independent of the covariates and normally distributed. As mentioned earlier more generally, the average partial effects are obtained directly from $P(y = 1|\mathbf{x})$, and, in fact, are given by $\beta_{cj}\phi(\mathbf{x}\boldsymbol{\beta}_c)$. As discussed in Wooldridge (2002, Chapter 15), $\beta_{cj}\phi(\mathbf{x}\boldsymbol{\beta}_c)$ can be larger or smaller in magnitude than the PEA $\beta_j\phi(\mathbf{x}\boldsymbol{\beta})$: $|\beta_{cj}| \leq |\beta_j|$ but $\phi(\mathbf{x}\boldsymbol{\beta}_c) \geq \phi(\mathbf{x}\boldsymbol{\beta})$.)

A related example is due to Hahn (2001), and is related to the nonidentification restuls of Chamberlain (1993). Suppose that x_{it} is a binary indicator (for example, a policy variable). Consider the unobserved effects probit model

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = \Phi(\beta x_{it} + c_i), \tag{5.3}$$

As discussed by Hahn, β is not known to be identified in this model, even under conditional serial independence assumption *and* the random effects assumption $D(c_i|\mathbf{x}_i) = D(c_i)$. But the average partial effect, which in this case is an average treatment effect, is simply $\tau \equiv E[\Phi(\beta + c_i)] - E[\Phi(c_i)]$. By the general result cited earlier, τ is consistently estimated (in fact, unbiasedly estimated) by using a difference of means for the treated and untreated groups, for either time period. In fact, as discussed in Wooldridge (2005a), identification of the APE holds if we replace Φ with an unknown function G and allow $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. But the parameters are still not identified.

To summarize: the APE is identified for any function $G(\cdot)$ whether or not the conditional serial independence holds, even if we add separate year intercepts. But β is not identified under the strongest set of assumptions. This simple example suggests that perhaps our focus on parameters is wrong-headed.

We can establish identification of average partial effects much more generally. Assume only that the strict exogeneity assumption (2.1) holds along with $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. These two assumptions are sufficient to identify the APEs. To see why, note that the average structural function at time t can be written as

$$\mathrm{ASF}_t(\mathbf{x}_t) = \mathrm{E}_{\mathbf{c}_i}[m_t(\mathbf{x}_t,\mathbf{c}_i)] \ = \mathrm{E}_{\bar{\mathbf{x}}_i}\{\mathrm{E}[m_t(\mathbf{x}_t,\mathbf{c}_i)|\bar{\mathbf{x}}_i]\} \ \equiv \mathrm{E}_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t,\bar{\mathbf{x}}_i)], \tag{5.4}$$

where $r_t(\mathbf{x}_t,\bar{\mathbf{x}}_i) \equiv \mathrm{E}[r_t(\mathbf{x}_t,\mathbf{c}_i)|\bar{\mathbf{x}}_i]$. It follows that, given an estimator $\hat{r}_t(\cdot,\cdot)$ of the function $r_t(\cdot,\cdot)$, the ASF can be estimated as

$$\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1}\sum_{i=1}^{N}\hat{r}_t(\mathbf{x}_t,\bar{\mathbf{x}}_i), \tag{5.5}$$

and then we can take derivatives or changes with respect to the entries in $\mathbf{x}_t$. Notice that (5.4) holds without the strict exogeneity assumption (2.1) or the assumption $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. However, these assumptions come into play in our ability to estimate $r_t(\cdot,\cdot)$. If we combine (21) and (4.3) we have

$$\begin{aligned}\mathrm{E}(y_{it}|\mathbf{x}_i) &= \mathrm{E}[\mathrm{E}(y_{it}|\mathbf{x}_i,\mathbf{c}_i)|\mathbf{x}_i] = \mathrm{E}[m_t(\mathbf{x}_{it},\mathbf{c}_i)|\mathbf{x}_i] = \int m_t(\mathbf{x}_{it},\mathbf{c})\mathrm{d}F(\mathbf{c}|\mathbf{x}_i) \\ &= \int m_t(\mathbf{x}_{it},\mathbf{c})\mathrm{d}F(\mathbf{c}|\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i),\end{aligned} \tag{5.6}$$

where $F(\mathbf{c}|\mathbf{x}_i)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_i)$ (which can be a discrete, continuous, or mixed distribution), the second equality follows from (2.1), the fourth equality follows from assumption (4.3), and the last equality folllows from the definition of $r_t(\cdot,\cdot)$ Of course, because $\mathrm{E}(y_{it}|\mathbf{x}_i)$ depends only on $(\mathbf{x}_{it},\bar{\mathbf{x}}_i)$, we must have

$$\mathrm{E}(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i). \tag{5.7}$$

Further, $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ is assumed to have time variation, and so $\mathbf{x}_{it}$ and $\bar{\mathbf{x}}_i$ can be used as separate regressors even in a fully nonparametric setting.

Altonji and Matskin (2005).use this idea more generally, and focus on estimating the local average response. Wooldridge (2005a) used $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$ generally in the case $\mathbf{x}_{it}$ is discrete, in which case a full nonparametric analysis is easy. When

$$\mathrm{D}(\mathbf{c}_i|\mathbf{x}_i) = \mathrm{D}(\mathbf{c}_i|\mathbf{w}_i) \tag{5.8}$$

for $\mathbf{w}_i$ a function of $\mathbf{x}_i$, Altonji and Matzkin (2005) show that the LAR can be obtained as

$$\int \frac{\partial r_t(\mathbf{x}_t,\mathbf{w})}{\partial x_{tj}}\mathrm{d}K_t(\mathbf{w}|\mathbf{x}_t), \tag{5.9}$$

where $r(\mathbf{x}_t,\mathbf{w}) = \mathrm{E}(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t,\mathbf{w}_i = \mathbf{w})$ and $K_t(\mathbf{w}|\mathbf{x}_t)$ is the cdf of $\mathrm{D}(\mathbf{w}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. Altonji and Matskin demonstrate how to estimate the LAR based on nonparametric estimation of $\mathrm{E}(y_{it}|\mathbf{x}_{it},\mathbf{w}_i)$ followed by "local" averaging, that is, averaging $\partial r(y_{it}|\mathbf{x}_t,\mathbf{w}_i)/\partial x_{tj}$ over observations i with $\mathbf{x}_{it}$ "close" to $\mathbf{x}_t$.

This analysis demonstrates that APEs are nonparametrically identified under the conditional mean version of strict exogeneity, (2.1), and (5.8), at least for time-varying covariates if $\mathbf{w}_i$ is restricted in some way. In fact, we can identify the APEs for a single time period with just one year of data on y. We only need to obtain $\bar{\mathbf{x}}_i$ and, in effect, include it as a control. Of course, efficiency would be gained by assuming some stationarity across t and using a pooled method.

6. Dynamic Models

General models with only sequentially exogenous variables are difficult to deal with. Arellano and Carrasco (2003) consider probit models. Wooldridge (2000) suggests a strategy the requires modeling the dynamic distribution of the variables that are not strictly exogenous. Much more is known about models with lagged dependent variables and otherwise strictly exogenous variables. So, we start with a model for

$$D(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\ldots,\mathbf{z}_{i1},\mathbf{y}_{i0},\mathbf{c}_i),\ t = 1,\ldots,T, \tag{6.1}$$

which we assume also is $D(\mathbf{y}_{it}|\mathbf{z}_i,\mathbf{y}_{i,t-1},\ldots,\mathbf{y}_{i1},\mathbf{y}_{i0},\mathbf{c}_i)$ where $\mathbf{z}_i$ is the entire history $\{z_{it} : t = 1,\ldots,T\}$. This is the sense in which the z_{it} are strictly exogenous.

Suppose this model depends only on $(\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c}_i)$, so $f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta})$. The joint density of $(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ given $(\mathbf{y}_{i0},\mathbf{z}_i,\mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}). \tag{6.2}$$

The problem with using this for estimation is the presence of $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$. Approaches: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (incidental parameters problem, although recent research has attempted to reduce the asymptotic bias in the partial effects). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for c_i. (Available in some special cases covered below, but other restrictions are needed. And, generally, cannot estimate partial effects.). (iii) Find or, more practically, approximate $D(\mathbf{y}_{i0}|\mathbf{c}_i,z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. After integrating out c_i we obtain the density for $D(\mathbf{y}_{i0},\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{z}_i)$ and we can use MLE (conditional on z_i), (iv) Model $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$. After integrating out c_i we obtain the density for $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{y}_{i0},\mathbf{z}_i)$, and we can use MLE (conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$). As shown by Wooldridge (2005b), in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ that mixes with the density $f(\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{y}_0,\mathbf{z},\mathbf{c})$ to produce a log-likelihood that is in a common family and carried out by standard software.

If $m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})$ is the mean function $E(y_t|\mathbf{x}_t,\mathbf{c})$ for a scalar y_t, then average partial effects are easy to obtain. The average structural function is

$$ASF(\mathbf{x}_t) = E_{c_i}[m_t(\mathbf{x}_t,\mathbf{c}_i,\theta)] \;= E\left\{\left[\int m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})h(\mathbf{c}|y_{i0},\mathbf{z}_i,\boldsymbol{\gamma})d\mathbf{c}\right]\Big|y_{i0},\mathbf{z}_i\right\}. \tag{6.3}$$

The term inside the brackets, say $r_t(x_t,y_{i0},z_i,\theta,\gamma)$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, in fact. Generally, it can be simulated. In any case, $ASF(\mathbf{x}_t,\boldsymbol{\theta})$ is consistently estimated by

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{t=1}^{T} r_t(\mathbf{x}_t,y_{i0},\mathbf{z}_i,\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}).$$

Partial derivatives and differences with respect to elements of x_t (which, remember, can include y_{t-1}) can be computed. With large N and small T, the panel data bootstrap can be used for standard errors and inference.

7. Applications to Specific Models

We now turn to some common parametric models and highlight the difference between estimation partial effects at different values of the heterogeneity and estimating average partial effects. An analysis of Tobit models follows in a very similar way to those in the following two sections. See Wooldridge (2002, Chapter 16) and Honoré and Hu (2004).

7.1 Binary and “Fractional” Response Models

We start with the standard specification for the unobserved effects (UE) probit model, which is

$$P(y_{it} = 1|\mathbf{x}_{it},c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1,\ldots,T, \tag{7.1}$$

where $\mathbf{x}_{it}$ does not contain an overall intercept but would usually include time dummies. We cannot identify $\boldsymbol{\beta}$ or the APEs without further assumptions. The traditional RE probit models imposes a strong set of assumptions: strict exogeneity, conditional serial independence, and independence between c_i and $\mathbf{x}_i$ with c_i ~Normal(μ_c,σ_c^2). Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE (conditional on $\mathbf{x}_i$).

We can relax independence between c_i and $\mathbf{x}_i$ using the Chamberlain-Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim\text{Normal}(0,\sigma_a^2), \tag{7.2}$$

where the time average is often used to save on degrees of freedom. We can relax (7.2) and allow Chamberlain's (1980) more flexible device:

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1}\boldsymbol{\xi}_1 +\ldots+\mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i \tag{7.3}$$

Even when the $\boldsymbol{\xi}_r$ seem to be very different, the Mundlak restriction can deliver similar estimates of the other parameters and the APEs. (In the linear case, they both produce the usual FE estimator of $\boldsymbol{\beta}$.)

If we still assume conditional serial independence then all parameters are identified. We can estimate the mean of c_i as $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^N \bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^N \bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Of course, c_i is not generally normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. The approximation might get better as T gets large. In any case, we can plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated as

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^N \Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{7.4}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\text{ASF}}(\mathbf{x}_t)$ with respect to elements of $\mathbf{x}_t$ can be compared with fixed effects estimates from a linear model. Often, if we also average out across $\mathbf{x}_{it}$, the linear FE estimates are similar to the averaged effects.

As we discussed generally in Section 5, the APEs are defined without the conditional serial independence assumption. Without $D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,c_i) = \prod_{t=1}^T D(y_{it}|\mathbf{x}_{it},c_i)$, we can still estimate the scaled parameters because

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a), \tag{7.5}$$

and so pooled probit consistently estimates the scaled parametes. (Time dummies have been supressed for simplicity.) Now we have direct estimates of $\boldsymbol{\beta}_a$, ψ_a, and $\boldsymbol{\xi}_a$, and we insert those directly into (7.4).

Using pooled probit can be inefficient for estimating the scaled parameters, whereas the full MLE is efficient but not (evidently) robust to violation of the conditional serial independence assumption. It is possible to estimate the parameters more efficiently than pooled probit that is still consistent under the same set of assumptions. One possibility is minimum distance estimation. That is, estimate a separate models for each t, and then impose the restrictions using minimum distance methods. (This can be done with or without the Mundlak device.)

A different approach is to apply the so called "generalized estimating equations" (GEE) approach. Briefly, GEE applied to panel data is essentially weighted multivariate nonlinear least squares (WMNLS) with explicit recognition that the weighting matrix might not be the inverse of the conditional variance matrix. In most nonlinear panel data models, obtaining the actual matrix $Var(\mathbf{y}_i|\mathbf{x}_i)$ is difficult, if not impossible, because integrating out the heterogeneity does not deliver a closed form. The GEE approach uses $Var(y_{it}|\mathbf{x}_i)$ implied by the specific distribution – in the probit case, we have the correct conditional variances,

$$Var(y_{it}|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)[1 - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)] \equiv v_{it}. \tag{7.6}$$

The "working" correlation matrix oftenusually specified as "exchangeable,"

$$Corr(e_{it}, e_{is}|\mathbf{x}_i) \text{ “ = ”} \rho, \tag{7.7}$$

where $e_{it} = [y_{it} - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)]v_{it}^{1/2}$ is the standardized error. Or, each pair (t,s) is allowed to have its own correlation but which is assumed to be independent of $\mathbf{x}_i$ ("unstructured"). The conditional correlation $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ is not constant, but that is the working assumption. The hope is to improve efficiency over the pooled probit estimator while maintaining the robustness of the pooled estimator. (The full RE probit estimator is not robust to serial dependence.) A robust sandwich matrix is easily computed provided the conditional mean function (in this case, response probability) is correctly specified.

Because the Bernoulli log-likelihood is in the linear exponential family (LEF), exactly the same methods can be applied if $0 \leq y_{it} \leq 1$ – that is, y_{it} is a "fractional" response – but where the model is for the conditional mean: $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$. Pooled "probit" or minimum distance estimation or GEE can be used. Now, however, we must make inference robust to $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ not having the probit form. (There are cases where $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ is proportional to (7.6), and so it still makes sense to use the probit quasi-log-likelihood. Pooled nonlinear regression is another possibility or weighted multivariate nonlinear regression are also possible and a special case of GEE.)

A more radical suggestion, but in the spirit of Altonji and Matzkin (2005) and Wooldridge (2005a), is to just use a flexible model for $\mathrm{E}(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i)$ directly. For example, if y_{it} is binary, or a fractional response, $0 \leq y_{it} \leq 1$, we might just specify a flexible parametric model, such as

$$E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.8}$$

or the “heteroskedastic probit” model

$$E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = \Phi[(\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})\exp(-\bar{\mathbf{x}}_i\boldsymbol{\eta})]. \tag{7.9}$$

If we write either of these functions as $r_t(\mathbf{x}_t,\bar{\mathbf{x}})$ then the average structural function is estimated as $\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1}\sum_{i=1}^{N}\hat{r}_t(\mathbf{x}_t,\bar{\mathbf{x}}_i)$, where the “^” indicates that we have substituted in the parameter estimates. We can let all parameters depend on t, or we can estimate the parameters separately for each t and then use minimum distance estimation to impose the parameter restrictions. The justification for using, say, (7.8) is that we are interested in the average partial effects, and how parameters appear is really not the issue. Even though (7.8) cannot be derived from $E(y_{it}|\mathbf{x}_{it},c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$ or any other standard model, there is nothing sacred about this formulation. In fact, it is fairly simplistic. We can view (7.8) as the approximation to the true $E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i)$ obtained after integrating c_i out of the unknown function $m(\mathbf{x}_t,c_i)$. (We could formalize this process by using series estimation, as in Newey (1988), where the number of terms is allowed to grow with N.) This is the same argument used by, say, Angrist (2001) in justifying linear models for limited dependent variables when the focus on primarily on average effects.

The argument is essentially unchanged if we replace $\bar{\mathbf{x}}_i$ with other statistics $\mathbf{w}_i$. For example, we might run, for each i, the regression $\mathbf{x}_{it}$ on $1,t$, $t = 1,\ldots,T$ and use the intercept and slope (on the time trend) as the elements of $\mathbf{w}_i$. Or, we can use sample variances and covariances for each i, along with the sample mean. Or, we can use initial values and average growth rates. The key condition is $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\mathbf{w}_i)$, and then we need sufficient variation in $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ not explained by $\mathbf{w}_i$ for identification. (Naturally, as we expand $\mathbf{w}_i$, the number of time periods required generally increases.)

Of course, once we just view (7.8) as an approximation, we can are justified in using the logistic function, say

$$E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = \Lambda[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.10}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$, which, again, can be applied to binary or fractional responses. The focus on partial effects that average out the heterogeneity can be liberating in that it means the step of specifying $E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$ is largely superfluous, and, in fact, can get in the way of pursuing a suitably flexible analysis. On the other hand, if we start with, say, a "structural" model such as $P(y_{i1} = 1|\mathbf{x}_i,\mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$, which is a heterogeneous index model, then we cannot derive equations such as (7.8) or (7.9), even under the strong assumption that $\mathbf{c}_i$ is independent of $\mathbf{x}_i$ and multivariate normal. If we imposed the Chamberlain device for the elements of $\mathbf{c}_i$ we can get expressions "close" to a combination of (7.8) and (7.9). Whether one is willing to simply estimate relative simple models such as (7.8) in order to estimate APEs depends on one's taste for bypassing more traditional formulations.

If we start with the logit formulation

$$P(y_{it} = 1|\mathbf{x}_{it},c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{7.11}$$

then we can estimate the parameters, $\boldsymbol{\beta}$ without restricting $D(c_i|\mathbf{x}_i)$ in any way, but we must add the conditional independence assumption. (No one has been able to show that, unlike in the linear model, or the Poisson model covered below, that the MLE that conditions on the number of successes $n_i = \sum_{t=1}^{T} y_{it}$ is robust to serial dependence. It appears not to be. Plus, the binary nature of y_{it} appears to be critical, so the conditional MLE cannot be applied to fractional responses even under serial independence.) Because we have not restricted $D(c_i|\mathbf{x}_i)$ in any way, it appears that we cannot estimate average partial effects. As commonly happens in nonlinear models, if we relax assumptions about the distribution of heterogeneity, we lose the ability to estimate partial effects. We can estimate the effects of the covariates on the log-odds ratio, and relative partial effects of continuous variables. But for partial effects themselves, we do not have sensible values to plug in for c, and we cannot average across its distribution.

The following table summarizes the features of various approaches to estimating binary response unobserved effects models.

Model, Estimation Method	$P(y_{it} = 1\|x_{it}, c_i)$	Restricts $D(c_i\|x_i)$?	Idiosyncratic Serial	PEs	APEs?
	Bounded in (0,1)?		Dependence?	at $E(c_i)$?	
RE Probit, MLE	Yes	Yes (indep, normal)	No	Yes	Yes
RE Probit, Pooled MLE	Yes	Yes (indep, normal)	Yes	No	Yes
RE Probit, GEE	Yes	Yes (indep, normal)	Yes	No	Yes
CRE Probit, MLE	Yes	Yes (lin. mean, normal)	No	Yes	Yes
CRE Probit, Pooled MLE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
CRE Probit, GEE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
LPM, Within	No	No	Yes	Yes	Yes
FE Logit, MLE	Yes	No	No	No	No

As an example, we apply several of the methods to women's labor force participation data, used by Chay and Hyslop (2001), where the data are for five time periods spaced four months apart. The results are summarized in the following table. The standard errors for the APEs were obtained with 500 bootstrap replications. The time-varying explanatory variables are log of husband's income and number of children, along with a full set of time period dummies. (The time-constant variables race, education, and age are also included in columns (2), (3), and (4).)

	(1)	(2)		(3)		(4)		(5)
Model	Linear	Probit		CRE Probit		CRE Probit		FE Logit
Estimation Method	Fixed Effects	Pooled MLE		Pooled MLE		MLE		MLE
	Coefficient	Coefficient	APE	Coefficient	APE	Coefficient	APE	Coefficient
kids	−.0389	−.199	−.0660	−.117	−.0389	−.317	−.0403	−.644
	(.0092)	(.015)	(.0048)	(.027)	(.0085)	(.062)	(.0104)	(.125)
lhinc	−.0089	−.211	−.0701	−.029	−.0095	−.078	−.0099	−.184
	(.0046)	(.024)	(.0079)	(.014)	(.0048)	(.041)	(.0055)	(.083)
$\overline{kids}$	—	—	—	−.086	—	−.210	—	—
	—	—	—	(.031)	—	(.071)	—	—
$\overline{lhinc}$	—	—	—	−.250	—	−.646	—	—
	—	—	—	(.035)	—	(.079)	—	—
$(1+\hat{\sigma}_a^2)^{-1/2}$	—	—		—		.387		—
Log Likelihood	—	−16,556.67		−16,516.44		−8,990.09		−2,003.42
Number of Women	5,663	5,663		5,663		5,663		1,055

Generally, CMLE approaches are fragile to changes in the specification. For example, a natural extension is

$$P(y_{it}=1|\mathbf{x}_{it},\mathbf{c}_i)=\Lambda(a_i+\mathbf{x}_{it}\mathbf{b}_i), \tag{7.12}$$

where $\mathbf{b}_i$ is a vector of heterogeneous slopes with $\boldsymbol{\beta}\equiv E(\mathbf{b}_i)$; let $\alpha\equiv E(a_i)$. This extension of the standard unobserved effects logit model raises several issues. First, what do we want to estimate? Perhaps the partial effects at the mean values of the heterogeneity. But the APEs, or local average effects, are probably of more interest.

Nothing seems to be known about what the logit CMLE would estimate if applied to (7.12), where we assume $\boldsymbol{\beta}=\mathbf{b}_i$. On the other hand, if, say, $D(\mathbf{c}_i|\mathbf{x}_i)=D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$, a flexible binary response model with covariates $(\mathbf{x}_{it},\bar{\mathbf{x}}_i)$ (and allowing sufficiently for changes over time) identifies the APEs – without the conditional serial independence assumption. The same is true of the extension to time-varying factor loads, $P(y_{it}=1|\mathbf{x}_{it},\mathbf{c}_i)=\Lambda(\theta_t+\mathbf{x}_{it}\boldsymbol{\beta}+\eta_t c_i)$.

There are methods that allow estimation, up to scale, of the coefficients without even specifying the distribution of u_{it} in

$$y_{it}=1[\mathbf{x}_{it}\boldsymbol{\beta}+c_i+u_{it}\geq 0]. \tag{7.13}$$

under strict exogeneity.conditional on c_i. Arellano and Honoré (2001) survey methods, including variations on Manski's maximum score estimator.

Estimation of parameters and APEs is much more difficult even in simple dynamic models. Consider

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, c_i) = P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i),\ t = 1, \ldots, T,$$

which combines correct dynamic specification with strict exogeneity of $\{z_{it}\}$. For a dynamic probit model

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{7.14}$$

Treating the c_i as parameters to estimate causes inconsistency in β and ρ (although there is recent work by Woutersen and Fernández-Val that shows how to make the asymptotic bias of order $1/T^2$; see the next section). A simple analysis is available if we specify

$$c_i|\mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i\xi, \sigma_a^2) \tag{7.15}$$

Then

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{7.16}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$. Because a_i is independent of $(y_{i0}, \mathbf{z}_i)$, it turns out we can use standard random effects probit software, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. Easily get the average partial effects, too:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{7.17}$$

and take differences or derivatives with respect to elements of $(\mathbf{z}_t, y_{t-1})$. As before, the coefficients are multiplied by $(1 + \hat{\sigma}_a^2)^{-1/2}$. Of course, both the structural model and model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $Var(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic).

We apply this method to the Chay and Hyslop data and estimate a model for P$(lfp_{it} = 1|kids_{it}, lhinc_{it}, lfp_{i,t-1}, c_i)$, where one lag of labor force participation is assumed to suffice for the dynamics and $\{(kids_{it}, lhinc_{it}) : t = 1, \ldots, T\}$ is assumed to be strictly exogenous conditional on c_i. Also, we include the time-constant variables *educ*, *black*, *age*, and age^2 and a full set of time-period dummies. (We start with five periods and lose one with the lag. Therefore, we estimate the model using four years of data.) We include among the regressors the initial value, lfp_{i0}, $kids_{i1}$ through $kids_{i4}$, and $lhinc_{i1}$ through $lhinc_{i4}$. Estimating the model by RE probit gives $\hat{\rho} = 1.541$ (se = .067), and so, even after controlling for unobserved heterogeneity, there is strong evidence of state dependence. But to obtain the size of the effect, we compute the APE for lfp_{t-1}. The calculation involves averaging $\Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a) - \Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a)$ across all t and i; we must be sure to scale the original coefficients by $(1 + \hat{\sigma}_a^2)^{-1/2}$, where, in this application, $\hat{\sigma}_a^2 = 1.103$. The APE estimated from this method is about .259. In other words, averaged across all women and all time periods, the probability of being in the labor force at time t is about .26 higher if the women was in the labor force at time $t-1$ than if she was not. This estimate controls for unobserved heterogeneity, number of young children, husband's income, and the woman's education, race, and age.

It is instructive to compare the APE with the estimate of a dynamic probit model that ignores c_i. In this case, we just use pooled probit of lfp_{it} on $1, kids_{it}, lhinc_{it}, lfp_{i,t-1} educ_i, black_i, age_i$, and age_i^2 and include a full set of period dummies. The coefficient on $lfp_{i,t-1}$ is 2.876 (se = $.027$), which is much higher than in the dynamic RE probit model. More importantly, the APE for state dependence is about $.837$, which is much higher than when heterogeneity is controlled for. Therefore, in this example, much of the persistence in labor force participation of married women is accounted for by the unobserved heterogeneity. There is still some state dependence, but its value is much smaller than a simple dynamic probit indicates.

Arellano and Carrasco (2003) use a different approach to estimate the parameters and APEs in dynamic binary response models with only sequentially exogenous variables. Thus, their method applies to models with lagged dependent variables, but also other models where there made be feedback from past shocks to future covariates. (Their assumptions essentially impose serial conditional serial independence.) Rather than impose an assumption such as (7.15), they use a different approximation. Let $v_{it} = c_i + u_{it}$ be the composed error in $y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]$. Then, in the context of a probit model, they assume

$$v_{it}|w_{it} \sim \text{Normal}(E(c_i|w_{it}), \sigma_t^2) \tag{7.18}$$

where $w_{it} = (x_{it}, y_{i,t-1}, x_{i,t-1}, \ldots, y_{i1}x_{i1})$. The mean $E(c_i|w_{it})$ is unrestricted (although, of course, they are linked across time by interacted expectations because $w_{it} \subset w_{i,t+1}$), but the shape of the distribution is assumed to be the same across t. Arellano and Carrasco discuss identification and estimation, and extensions to models with time-varying factor loads.

Honoré and Kyriazidou (2000) extend an idea of Chamberlain's and show how to estimate $\boldsymbol{\delta}$ and ρ in a logit model without distributional assumptions for c_i. They find conditional probabilities that do not depend on c_i but still depend on $\boldsymbol{\delta}$ and ρ. However, in the case with four time periods, $t = 0, 1, 2$, and 3, the conditioning that removes c_i requires $z_{i2} = z_{i3}$. HK show how to use a local version of this condition to consistenty estimate the parameters. The estimator is also asymptotically normal, but converges more slowly than the usual $\sqrt{N}$-rate.

The condition that $z_{i2} - z_{i3}$ have a distribution with support around zero rules out aggregate year dummies or even linear time trends. Plus, using only observations with $z_{i2} - z_{i3}$ in a neighborhood of zero results in much lost data. Finally, estimates of partial effects or average partial effects are not available.

While semiparametric approaches can be valuable to comparing parameter estimates with more parametric approaches, such comparisons have limitations. For example, the coefficients on y_{t-1} in the dynamic logit model and the dynamic probit model are comparable only in sign; we cannot take the derivative with respect to y_{t-1} because it is discrete. Because we do not know where the evaluate the partial effects – that is, the values of c to plug in, or average out across the distribution of c_i, we cannot compare the magnitudes with CRC approaches. We can compare the relative effects on the continuous elements in $\mathbf{z}_t$ based on partial derivatives. But even here, if we find a difference between semiparametric and parametric methods, is it because aggregate time effects were excluded in the semiparametric estimation or because the model of $D(c_i | y_{i0}, \mathbf{z}_i)$ was misspecified? Currently, we have no good ways of deciding. (Recently, Li and Zheng (2006) use Bayesian methods to estimate a dynamic Tobit model with unobserved heterogeneity, where they distribution of unosberved heterogeneity is an infinite mixture of normals. They find that all of the average partial effects are very similar to those obtained from the much simpler specification in (7.15).)

Honoré and Lewbel (2002) show how to estimate β in the model

$$y_{it} = 1[v_{it} + x_{it}\beta + c_i + u_{it} \geq 0] \tag{7.19}$$

without distributional assumptions on $c_i + u_{it}$. The special continuous explanatory variable v_{it}, which need not be time varying, is assumed to appear in the equation (and its coefficient is normalized to one). More importantly, v_{it} is assumed to satisfy $D(c_i + u_{it} | v_{it}, x_{it}, z_i) = D(c_i + u_{it} | x_{it}, z_i)$, which is a conditional independence assumption. The vector z_i is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, z_i could be functions of variables determined prior to the earliest time period.) The most likely scenario is when v_{it} is randomized and therefore independent of (x_{it}, z_i, e_{it}), where $e_{it} = c_i + u_{it}$. It seems unlikely to hold if v_{it} is related to past outcomes on y_{it}. The estimator derived by Honoré and Lewbel is $\sqrt{N}$-asymptotically normal, and fairly easy to compute; it requires estimation of the density of v_{it} given (x_{it}, z_i) and then a simple IV estimation.

Honoré and Tamer (2006) have recently shown how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model; these are covered in the notes on partial identification.

7.2 Count and Other Multiplicative Models

Several options are available for models with conditional means multiplicative in the heterogeneity. The most common is

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.20}$$

where $c_i \geq 0$ is the unobserved effect and x_{it} would incude a full set of year dummies in most cases. First consider estimation under strict exogeneity (conditional on c_i):

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i). \tag{7.21}$$

If we add independence between c_i and x_i – a random effects approach – then, using $E(c_i) = 1$ as a normalization,

$$E(y_{it}|\mathbf{x}_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}), \tag{7.22}$$

and various estimation methods can be used to account for the serial dependence in $\{y_{it}\}$ if only x_i is conditioned on. (Serial correlation is certainly present because of c_i, but it could be present due to idiosyncratic shocks, too.) Regardless of the actual distribution of y_{it}, or even its nature – other than $y_{it} \geq 0$ – the pooled Poisson quasi-MLE is consistent for β under (7.22) but likely very inefficient; robust inference is straightforward with small T and large N.

Random effects Poisson requires that $D(y_{it}|\mathbf{x}_i, c_i)$ has a Poisson distribution with mean (7.20), and maintains the conditional independence assumption,

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it}, c_i),$$

along with a specific distribution for c_i – usually a Gamma distribution with unit mean. Unfortunately, like RE probit, the full MLE has no known robustness properties. The Poisson distribution needs to hold along with the other assumptions. A generalized estimating approach is available, too. If the Poisson quasi-likelihood is used, the GEE estimator is fully robust provided the mean is correctly specified. One can use an exchangeable, or at least constant, working correlation matrix. See Wooldridge (2002, Chapter 19).

A CRE model can be allowed by writing $c_i = \exp(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi})a_i$ where a_i is independent of x_i

with unit mean. Then

$$E(y_{it}|\mathbf{x}_i) = \exp(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}) \tag{7.23}$$

and now the same methods described above can be applied but with $\bar{x}_i$ added as regressors. This approach identifies average partial effects. In fact, we could use Altonji and Matzkin (2005) and specify $E(c_i|x_i) = h(\bar{\mathbf{x}}_i)$ (say), and then estimate the semiparametric model $E(y_{it}|\mathbf{x}_i) = h(\bar{\mathbf{x}}_i)\exp(\mathbf{x}_{it}\boldsymbol{\beta})$. Other features of the series $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$, such as individual-specific trends or sample variances, can be added to $h(\bullet)$.

An important estimator that can be used under just

$$E(y_{it}|\mathbf{x}_i, c_i) = c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.24}$$

is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption. It is often called the fixed effects Poisson estimator, and, in fact, $\hat{\boldsymbol{\beta}}$ turns out to be identical to using pooled Poisson QMLE and treating the c_i as parameters to estimate. (A rare case, like the linear model, where this does not result in an incidental parameters problem.). It is easy to obtain fully robust inference, too (although it is not currently part of standard software, such as Stata). The fact that the quasi-likelihood is derived for a particular, discrete distribution appears to make people queasy about using it, but it is analogous to using the normal log-likelihood in the linear model: the resulting estimator, the usual FE estimator, is fully robust to nonnormality, heteroskedasticity, and serial correlation.

Estimation of models under sequential exogeneity has been studied by Chamberlain (1992) and Wooldridge (1997). In particular, they obtain moment conditions for models such as

$$E(y_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}, c_i) = c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta}). \tag{7.25}$$

Under this assumption, it can be shown that

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}] = 0, \tag{7.26}$$

and, because these moment conditions depend only on observed data and the parameter vector $\boldsymbol{\beta}$, GMM can be used to estimate $\boldsymbol{\beta}$, and fully robust inference is straightforward.

The moment conditions in (7.26) involve the differences $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$, and we saw for the linear model that, if elements of $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$ are persistent, IV and GMM estimators can be badly biased and imprecise. If we make more assumptions, models with lagged dependent variables and other regressors that are strictly exogenous can be handled using the conditional MLE approach in Section 6. Wooldridge (2005b) shows how a dynamic Poisson model with conditional Gamma heterogeneity can be easily estimated.

8. Estimating the Fixed Effects

It is well known that, except in special cases (linear and Poisson), treating the c_i as parameters to estimate leads to inconsistent estimates of the common parameters $\boldsymbol{\theta}$. But two questions arise. First, are there ways to adjust the "fixed effects" estimate of $\boldsymbol{\theta}$ to at least partially remove the bias? Second, could it be that estimates of the average partial effects, based generally on

$$N^{-1}\sum_{i=1}^{N}\frac{\partial m_t(\mathbf{x}_t,\hat{\boldsymbol{\theta}},\hat{\mathbf{c}}_i)}{\partial x_{tj}}, \tag{8.1}$$

where $m_t(\mathbf{x}_t,\boldsymbol{\theta},\mathbf{c}) = E(y_t|\mathbf{x}_t,\mathbf{c})$, are better behaved than the parameter estimates, and can their bias be removed? In the unobserved effects probit model, (8.1) becomes

$$N^{-1}\sum_{i=1}^{N}\hat{\beta}_j\phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}+\hat{c}_i), \tag{8.2}$$

which is easy to compute once $\hat{\boldsymbol{\beta}}$ and the $\hat{c}_i$ (N of them) have been obtained.

Hahn and Newey (2004) propose both jackknife and analytical bias corrections and show that they work well for the probit case. Generally, the jackknife procedure to remove the bias in $\hat{\boldsymbol{\theta}}$ is simple but can be computationally intensive. The idea is this. The estimator based on T time periods has probability limit that can be written as

$$\boldsymbol{\theta}_T = \boldsymbol{\theta} + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \tag{8.3}$$

for vectors $\mathbf{b}_1$ and $\mathbf{b}_2$. Now, let $\hat{\boldsymbol{\theta}}_{(t)}$ denote the estimator that drops time period t. Then, assuming stability across t, the plim of $\hat{\boldsymbol{\theta}}_{(t)}$ is

$$\boldsymbol{\theta}_{(t)} = \boldsymbol{\theta} + \mathbf{b}_1/(T-1) + \mathbf{b}_2/(T-1)^2 + O(T^{-3}). \tag{8.4}$$

It follows that

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\,(T\hat{\boldsymbol{\theta}} - (T-1)\hat{\boldsymbol{\theta}}_{(t)}) &= (T\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/T) - [(T-1)\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/(T-1)] + O(T^{-3}) \\ &= \boldsymbol{\theta} - \mathbf{b}_2/[T(T-1)] + O(T^{-3}) = \boldsymbol{\theta} + O(T^{-2}).\end{aligned} \tag{8.5}$$

If, for given heterogeneity c_i, the data are independent and identically distributed, then (8.5) holds for all leave-one-time-period-out estimators, so we use the average of all such estimators in computing the panel jackknife estimator:

$$\tilde{\boldsymbol{\theta}} = T\hat{\boldsymbol{\theta}} - (T-1)T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\theta}}_{(t)}. \tag{8.6}$$

From the argument above, theasymptotic bias of $\tilde{\boldsymbol{\theta}}$ is on the order of T^{-2}.

Unfortunately, there are some practical limitations to the jackknife procedure, as well as to the analytical corrections derived by Hahn and Newey. First, aggregate time effects are not allowed, and they would be very difficult to include because the analysis is with $T \rightarrow \infty$. (In other words, they would introduce an incidental parameters problem in the time dimension as well as cross section dimension.) Generally, heterogeneity in the distributions across t changes the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ when a time period is dropped, and so the simple transformation in (8.5) does not remove the bias terms. Second, Hahn and Newey assume independence across t conditional on c_i. It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, as noted by Hahn and Keursteiner, applying the "leave-one-out" method to dynamic models is problematical because the $\mathbf{b}_1$ and $\mathbf{b}_2$ in (8.4) would depend on t so, again, the transformation in (8.5) will not eliminate the $\mathbf{b}_1$ term.

Recently, Dhaene, Jochmans, and Thuysbaert (2006) propose a modification of the Hahn-Newey procedure that appears promising for dynamic models. In the simplest case, in addition to the "fixed effects" estimator using all time periods, they obtain estimators for two subperiods: one uses the earlier time periods, one uses later time periods, and they have some overlap (which is small as T gets large). Unfortunately, the procedure still requires stationarity and rules out aggregate time effects.

For the probit model, Fernández-Val (2007) studies the properties of estimators and average partial effects and allows time series dependence in the strictly exogenous regressors. Interestingly, in the probit model with exogenous regressors under the conditional independence assumption, the estimates of the APEs based on the "fixed" effects estimator has bias of order T^{-2} in the case that there is no heterogeneity. Unfortunately, these findings do not carry over to models with lagged dependent variables, and the bias corrections in that case are difficult to implement (and still do not allow for time heterogeneity).

The correlated random effects estimators restrict $D(c_i|\mathbf{x}_i)$ in some way, although the recent work by Altonji and Matzkin (2005) shows how those restrictions can be made reasonable. The approach generally identifies the APEs, and even the local average effects, and does not rule out aggregate time effects or arbitrary conditional serial dependence.

References

(To be added.)