# What's New in Econometrics?
# Lecture 4
# Nonlinear Panel Data Models

Jeff Wooldridge
NBER Summer Institute, 2007

1. Basic Issues and Quantities of Interest

2. Exogeneity Assumptions

3. Conditional Independence

4. Assumptions about the Unobserved Heterogeneity

5. Nonparametric Identification of Average Partial Effects

6. Dynamic Models

7. Applications to Specific Models

8. Estimating the Fixed Effects

## 1. Basic Issues and Quantities of Interest

● Let $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$ be a random draw from the cross section. Typically interested in

$$D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \tag{1}$$

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? If $x_{tj}$ is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}, \tag{2}$$

or discrete changes. How do we account for unobserved $\mathbf{c}_i$? If we know enough about the distribution of $\mathbf{c}_i$ 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),*

$$PEA_j(\mathbf{x}_t) = \theta_j(\mathbf{x}_t, \boldsymbol{\mu}_{\mathbf{c}}). \tag{3}$$

Of course, we need to estimate the function $m_t$ and $\boldsymbol{\mu}_{\mathbf{c}}$. We might be able to insert different quantiles, or a certain number of standard deviations from the mean.

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

$$APE(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)]. \tag{4}$$

The difference between (3) and (4) can be nontrivial. In some leading cases, (4) is identified while (3) is not. (4) is closely related to the notion of the average structural function (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$ASF(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \tag{5}$$

● Passing the derivative through the expectation in

(5) gives the APE.

- How do APEs relate to parameters? Suppose

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \qquad (6)$$

where, say, $G(\bullet)$ is strictly increasing and continuously differentiable. Then

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \qquad (7)$$

where $g(\bullet)$ is the derivative of $G(\bullet)$. Then estimating $\beta_j$ means we can sign of the partial effect, and the relative effects of any two continuous variables. Even if $G(\bullet)$ is specified, the magnitude of effects cannot be estimated without making assumptions about the distribution of $c_i$

- Altonji and Matzkin (2005) define 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

$$LAR_j(\mathbf{x}_t) = \int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t), \tag{8}$$

where $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. "Local" because it averages out the heterogeneity for the slice of the population described by the vector $\mathbf{x}_t$. The APE is a "global" average response."

- Definitions of partial effects do not depend on whether $\mathbf{x}_t$ is correlated with $\mathbf{c}$. Of course, whether and how we estimate them certainly does.

**2. Exogeneity Assumptions**

- As in linear case, cannot get by with just specifying a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$.
- The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \tag{9}$$

Chamberlain (1984) labeled (9) *strict exogeneity* *conditional on the unobserved effects* $\mathbf{c}_i$.

Conditional mean version:

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \quad (10)$$

● The sequential exogeneity assumption is

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{it},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \quad (11)$$

Unfortunately, it is much more difficult to allow sequential exogeneity in in nonlinear models.

● Neither (9) nor (10) 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 $y_{it}$. (Later in control function estimation.)

**3. Conditional Independence**

● In linear models, serial dependence of idiosyncratic shocks is easily dealt with, either by robust inference or GLS extensions of FE and FD. With strictly exogenous covariates, never results in biased estimation, even if it is ignored or improperly model. The situation is different with nonlinear models estimated by MLE.

- The conditional independence assumption is

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i) \tag{12}$$

(where we also impose strict exogeneity). In a parametric context, the CI 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 correlated random effects frameworks, CI plays a critical role in being able to estimate the "structural" parameters and the parameters in distribution the of $\mathbf{c}_i$ (and therefore, PAEs). In a broad class of models, CI plays no role in estimating APEs.

## 4. Assumptions about the Unobserved Heterogeneity

### Random Effects

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{13}$$

Under (13), the APEs are nonparametrically identified from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{14}$$

- In some leading cases (RE probit and RE Tobit with heterogeneity normally distributed), if we want PEs for different values of **c**, we must assume more: strict exogeneity, conditional independence, and (13) with a parametric distribution for $D(\mathbf{c}_i)$.

**Correlated Random Effects**

A CRE framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but 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 Chamberlain (1980,1982), and much work since. Can allow $D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$ to depend in a "nonexchangeable" manner. (Chamberlain's CRE probit and Tobit models.) Distributional assumptions that lead to simple estimation – homoskedastic normal with a linear conditional mean — are restrictive.

- Possible to drop parametric assumptions with

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i), \tag{15}$$

without restricting $\mathrm{D}(c_i|\bar{\mathbf{x}}_i)$.

- As $T$ gets larger, can allow $\mathbf{c}_i$ to be correlated with features of the covariates other than just the time average. Altonji and Matzkin (2005) allow for $\bar{\mathbf{x}}_i$ in equation (15) to be replaced by other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as sample variances and covariance. Non-exchangeable functions, such as unit-specific trends, can be used, too. Generally, assume

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\mathbf{w}_i). \tag{16}$$

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 $\mathbf{w}_i$.

- Altonji and Matzkin use exchangeability and other restrictions, such as monotonicity

**Fixed Effects**

The label "fixed effects" is used in different ways by different researchers. One view: $\mathbf{c}_i,\ i = 1, \ldots, N$

are parameters to be estimated. Usually leads to an "incidental parameters problem" (which attentuates with large $T$).

- A second meaning of "fixed effects" is that $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted and we look for objective functions that do not depend on $\mathbf{c}_i$ but still identify the population parameters. Leads to "conditional maximum likelihood" if we can find a "sufficient statistic" such that

$$D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{s}_i). \tag{17}$$

- The CI assumption is usually maintained.

## 5. Nonparametric Identification of Average Partial Effects

- Identification of PAEs can fail even under a strong set of parametric assumptions. In the probit model

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\boldsymbol{\beta} + c), \tag{18}$$

the PE for a continuous variable $x_j$ is $\beta_j \phi(\mathbf{x}\boldsymbol{\beta} + c)$. The PAE at $\mu_c = E(c) = 0$ is $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$. Suppose $c|\mathbf{x}$ ~Normal$(0, \sigma_c^2)$. Then

$$P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}), \tag{19}$$

so only the scaled parameter vector $\boldsymbol{\beta}_c \equiv \boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}$ is identified; $\boldsymbol{\beta}$ and $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$ are not identified.

- The APE is identified from $P(y = 1|\mathbf{x})$, and is given by $\beta_{cj}\phi(\mathbf{x}\boldsymbol{\beta}_c)$. (Attenuation bias?)
- Panel data example due to Hahn (2001): $x_{it}$ is a binary indicator and

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = \Phi(\beta x_{it} + c_i), t = 1, 2. \tag{20}$$

$\beta$ is not known to be identified in this model, even under conditional independence *and* the random effects assumption $D(c_i|\mathbf{x}_i) = D(c_i)$. But the APE is $\tau \equiv E[\Phi(\beta + c_i)] - E[\Phi(c_i)]$ and is identified by a difference of means for the treated and untreated groups, for either time period.

● As shown in Wooldridge (2005a), identification of the APE holds if we replace $\Phi$ with an unknown function $G$ and allow $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$.

● Are we focusing too much on parameters? In many cases, yes, but not always so clear cut. From Wooldridge (2005c): $y = 1[\mathbf{x}\boldsymbol{\beta} + u > 0]$ where $u|\mathbf{x}$ ~Normal$(0, \exp(2\mathbf{x}\boldsymbol{\delta}))$ ("heteroskedastic probit"). $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ estimable by MLE. The APE for $x_j$ is *not* obtained by differentiating $P(y = 1|\mathbf{x}) = \Phi[\exp(-\mathbf{x}\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$ with respect to $x_j$, which can have a different sign from $\beta_j$. Instead, for given $\mathbf{x}$, it is consistently estimated as

$$\widehat{APE}_j(\mathbf{x}) = \hat{\beta}_j \left\{ N^{-1} \sum_{i=1}^{N} \phi[\exp(-\mathbf{x}_i \hat{\boldsymbol{\delta}}) \mathbf{x} \hat{\boldsymbol{\beta}}] \right\},$$

which always has the same sign as $\hat{\beta}_j$.

- We can establish identification of APEs in panel data applications very under strict exogeneity along with $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. These two assumptions identify the APEs. Write the average structural function at time $t$ as

$$\begin{aligned} \mathrm{ASF}_t(\mathbf{x}_t) &= \mathrm{E}_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] \\ &= \mathrm{E}_{\bar{\mathbf{x}}_i}\{\mathrm{E}[m_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]\} \\ &\equiv \mathrm{E}_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)], \end{aligned} \tag{21}$$

Given a consistent estimator of $\hat{r}_t(\cdot,\cdot)$, the ASF can be estimated as

$$\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1} \sum_{i=1}^{N} \hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i).\,, \tag{22}$$

● Equation (21) holds without strict exogeneity $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. But these assumptions allow us to estimate estimate $r_t(\cdot,\cdot)$:

$$\begin{aligned} E(y_{it}|\mathbf{x}_i) &= E[E(y_{it}|\mathbf{x}_i,\mathbf{c}_i)|\mathbf{x}_i] = E[m_t(\mathbf{x}_{it},\mathbf{c}_i)|\mathbf{x}_i] \\ &= \int m_t(\mathbf{x}_{it},\mathbf{c})dF(\mathbf{c}|\mathbf{x}_i) \\ &= \int m_t(\mathbf{x}_{it},\mathbf{c})dF(\mathbf{c}|\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i), \end{aligned} \tag{23}$$

where $F(\mathbf{c}|\mathbf{x}_i)$ denotes the cdf of $D(\mathbf{c}_i|\mathbf{x}_i)$ Because $E(y_{it}|\mathbf{x}_i)$ depends only on $(\mathbf{x}_{it},\bar{\mathbf{x}}_i)$, we must have

$$E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i), \tag{24}$$

and $r_t(\cdot,\cdot)$ is identified with sufficient time variation in $\mathbf{x}_{it}$.

## 6. Dynamic Models

● Nonlinear models with only sequentially exogenous variables are difficult to deal with. More is known about models with lagged dependent variables and otherwise strictly exogenous variables:

$$D(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \ldots, \mathbf{z}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i),\ t = 1, \ldots, T, \tag{25}$$

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

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}). \tag{26}$$

- How do we deal with $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$? Approaches: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (incidental parameters problem). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for $c_i$ (available in some special cases). Generally, cannot estimate partial effects.). (iii) Approximate $D(\mathbf{y}_{i0}|\mathbf{c}_i, z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. Leads to $D(\mathbf{y}_{i0}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{z}_i)$ and MLE conditional on $\mathbf{z}_i$. (iv) Model $D(\mathbf{c}_i|\mathbf{y}_{i0}, \mathbf{z}_i)$. Leads to $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{y}_{i0}, \mathbf{z}_i)$ and MLE conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$. Wooldridge (2005b) shows this can be computationally simple for popular models.

- 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 $y_t$, the APEs are easy to obtain.

## 7. Applications to Specific Models

### Binary and Fractional Response

- Unobserved effects (UE) probit model:

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1, \ldots, T. \tag{27}$$

Assume strict exogeneity (as always, conditional on $c_i$) and use Chamberlain-Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i, a_i | \mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2). \tag{28}$$

If we still assume conditional serial independence then all parameters are identified and MLE (RE probit) can be used. $\hat{\mu}_c = \hat{\psi} + \left(N^{-1} \sum_{i=1}^{N} \bar{\mathbf{x}}_i\right) \hat{\boldsymbol{\xi}}$ and $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}' \left(N^{-1} \sum_{i=1}^{N} \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i\right) \hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. $c_i$ is not generally normally distributed unless $\bar{\mathbf{x}}_i \boldsymbol{\xi}$ is. But can evaluate PEs at, say, $\hat{\mu}_c \pm k \hat{\sigma}_c$.

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

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_t \hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_a) \tag{29}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}} / (1 + \hat{\sigma}_a^2)^{1/2}$.

- APEs are identified without the conditional serial independence assumption. Use the marginal probabilities to estimate scaled coefficients:

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a). \tag{30}$$

(Time dummies have been supressed for simplicity.)

- Can used pooled probit or minimum distance or "generalized estimating equations."
- 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: $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$. Full MLE difficult.
- A more radical suggestion, but in the spirit of Altonji and Matzkin (2005), is to just use a flexible model for $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly, say,

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

Just average out over $\bar{\mathbf{x}}_i$ to get APEs.

- Can use same idea with logit. But, if we have a binary response, start with

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{31}$$

and assume conditional independence assumption, we can estimate $\boldsymbol{\beta}$ without restricting $D(c_i|\mathbf{x}_i)$.

- Because we have not restricted $D(c_i|\mathbf{x}_i)$ in any way, it appears that we cannot estimate average partial effects. See table in notes for the tradeoffs in using CRE models and conditional MLE.

- Example from notes. Estimated APEs for number of small children on women's labor force participation: linear, $-.0389$ (.0092); probit (pooled), $-.0660$ (.0048); CRE probit (pooled) $-.0389$ (.0085); CRE probit (MLE), $-.0403$ (.104), FE logit, coefficient = $-.644$ (.125).

- What would CMLE logit estimate in the model

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(a_i + \mathbf{x}_{it}\mathbf{b}_i), \tag{32}$$

where $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$?

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

$$y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]. \tag{33}$$

under strict exogeneity.conditional on $c_i$. Arellano and Honoré (2001).

- Simple dynamic model:

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{34}$$

A simple analysis is available if we specify

$$c_i | \mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i \xi, \sigma_a^2) \tag{35}$$

Then

$$P(y_{it} = 1 | \mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i \boldsymbol{\xi} + a_i), \tag{36}$$

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:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{z}_t \hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i \hat{\boldsymbol{\xi}}_a), \tag{37}$$

Example in notes: dynamic labor force partication. The APE estimated from this method is about .259. If we ignore the heterogeneity, APE is .837.

- For estimating parameters, Honoré and Kyriazidou (2000) extend an idea of Chamberlain. With four 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. By design, cannot estimate magnitudes of effects.

**Count and Other Multiplicative Models**

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

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{38}$$

where $c_i \geq 0$. If we assume strict exogeneity,

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i), \tag{39}$$

a particular quasi-MLE is attractive as it does not restrict $D(y_{it}|\mathbf{x}_i, c_i)$, $D(c_i|\mathbf{x}_i)$, or serial dependence: the "fixed effects" Poisson estimator. It is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption. But it is fully robust, even if $y_{it}$ is not a count variable! It turns out that there is no incidental parameters problem in this case. Fully robust inference is easy to obtain (Wooldridge (1999)).

- Estimation under sequential exogeneity has been studied by Chamberlain (1992) and Wooldridge

(1997). In particular, they obtain moment conditions for models such as

$$E(y_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}). \quad (40)$$

Under this assumption, it can be shown that

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}] = 0, \quad (41)$$

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.

- Wooldridge (2005b) shows how a dynamic Poisson model with conditional Gamma heterogeneity can be easily estimated.

## 8. Estimating the Fixed Effects

- Except in special cases (linear and Poisson), treating the $c_i$ as parameters to estimate leads to inconsistent estimates of the population parameters $\boldsymbol{\theta}$. But 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 APEs, based on

$$N^{-1}\sum_{i=1}^{N}\frac{\partial m_t(\mathbf{x}_t,\hat{\boldsymbol{\theta}},\hat{\mathbf{c}}_i)}{\partial x_{tj}}, \tag{42}$$

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?

- Hahn and Newey (2004) propose both jackknife and analytical bias corrections and show that they work well for the probit case. The jackknife FE estimator is

$$\tilde{\theta} = T\hat{\theta} - (T-1)T^{-1}\sum_{t=1}^{T}\hat{\theta}_{(t)}, \tag{43}$$

where $\hat{\theta}$ is the FE estimate using all time periods and $\hat{\theta}_{(t)}$ is the estimate that drops time period $t$. The asymptotic bias of $\tilde{\theta}$ is on the order of $T^{-2}$.

- Practical limitations of the jackknife. First, aggregate time effects are not allowed, and they would be difficult to include because the analysis is with $T \to \infty$. Also, heterogeneity in the distributions across $t$ changes the bias terms and so (43) does not remove the bias. Hahn and Newey assume independence across $t$ conditional on $c_i$. Even relaxing this, the “leave-one-out” method does not apply to dynamic models.
- Fernández-Val (2007) shows that in a model with time series dependence in strictly exogenous

regressors, the APEs based on the fixed effects estimator have bias of order $T^{-2}$ in the case that there is no heterogeneity.