Nonlinear Panel Data Models

Manuel Arellano

CEMFI

September 2013

I. Introduction

- Nonlinear panel data models with individual effects are very common in economics.

Examples of nonlinear models

- *Discrete choice* models:

$$y_{it} = 1\left(x_{it}'\theta + \alpha_i + v_{it} > 0\right)$$

  e.g. labor force participation (Hyslop, 1999).

- *VAR models* of transmission of shocks:

$$\begin{aligned} y_{it} &= (\beta y_{it-1} + \alpha_{1i} + v_{it})\, d_{it} \\ d_{it} &= 1\left(\gamma d_{it-1} + \alpha_{2i} + \phi v_{it} + \varepsilon_{it}\right) \end{aligned}$$

  e.g. employment status and earnings (Altonji, Smith, and Vidangos, 2009).

Nonlinear examples (continued)

- *Distributional dynamics* in a location–scale model:

$$y_{it} = x_{it}'\theta_1 + \alpha_{1i} + \sigma\left(x_{it}'\theta_2, \alpha_{2i}\right)\varepsilon_{it}$$

  e.g. earnings dynamics (Meghir and Pistaferri, 2004; Hospido, 2012).

- A semiparametric generalization of the above is the quantile model

$$y_{it} = x_{it}'\beta\left(u_{it}\right) + \alpha_i\gamma\left(u_{it}\right)$$

  where u_{it} is the rank of the error v_{it}, so that

$$u_{it} \mid x_{i1}, ..., x_{iT}, \alpha_i \quad \sim \quad \mathcal{U}\left(0,1\right),$$

  and $\beta\left(u\right)$ and $\gamma\left(u\right)$ are nonparametric functions.

Nonlinear examples (continued)

- *Structural models* with unobserved heterogeneity
  e.g. schooling choice, search-matching models, production functions...

- Non-additive fixed effects may also arise in continuous response functions. An example is the following heterogeneous constant elasticity of substitution (CES) production function:

$$\log y_{it} = \lambda \log h_{it} + (1-\lambda) \log \left[\gamma x_{it}^{\sigma_i} + (1-\gamma) z_{it}^{\sigma_i} \right]^{1/\sigma_i} + \alpha_i + v_{it},$$

- This model allows for different degrees of complementarity between high-skill labor (h_{it}), low-skill labor (x_{it}), and capital equipment (z_{it}).

Remarks

Additive vs non-additive errors

- Linear panel ideas generalize easily to nonlinear models with additive errors. These include nonlinear WG:

$$y_{it} = g_t(x_{it}, \theta_0) + \alpha_{i0} + v_{it} \text{ where } E(v_{it}|x_i, \alpha_{i0}) = 0$$

  and nonlinear implicit structural equations (Euler equations, production functions):

$$\rho_t(w_{it}, \theta_0) = \alpha_{i0} + v_{it} \text{ where } E(v_{it}|z_i, \alpha_{i0}) = 0.$$

- For these models one can construct moment conditions that mimic the linear ones.
- Linear models with random coefficients generalize to nonlinear models that are linear in the random coefficients:

$$y_{it} = g_0(x_{it}, \theta) + g_1(x_{it}, \theta)' \alpha_i + v_{it}.$$

  This model was studied in Chamberlain (1992) and has been recently re-examined in Arellano & Bonhomme (2012) and Graham & Powell (2012).
- The situation is fundamentally different in the absence of additivity. A leading example is the binary choice model.

Remarks (continued)

Policy parameters (derivative effects)

- Effect on y of changing x from x_A to x_B. In linear models:

$$(x_B\theta_0 + \alpha_{i0} + v_{it}) - (x_A\theta_0 + \alpha_{i0} + v_{it}) = (x_B - x_A)\,\theta_0$$

- In binary choice the effect is individual-specific:

$$1\,(x_B\theta_0 + \alpha_{i0} + v_{it} \geq 0) - 1\,(x_A\theta_0 + \alpha_{i0} + v_{it} \geq 0)$$

  Letting F be the cdf of v, the average effect for a given α_{i0} is

$$F\,(x_B\theta_0 + \alpha_{i0}) - F\,(x_A\theta_0 + \alpha_{i0})$$

- The conclusion is that in nonlinear models derivative effects mix common and individual effects.

Average derivative effects

- A derivative version of the above is

$$\frac{\partial F\left(x\theta_0+\alpha_{i0}\right)}{\partial x}\mid_{x=x_A}$$

- We may wish to consider averages wrt α_{i0} using either the marginal density of α_{i0} (Chamberlain 1984):

$$\int\frac{\partial F\left(x\theta_0+\alpha_{i0}\right)}{\partial x}\mid_{x=x_A} dG\left(\alpha_{i0}\right)$$

  or the density of α_{i0} conditioned on $x=x_A$:

$$\int\frac{\partial F\left(x\theta_0+\alpha_{i0}\right)}{\partial x}\mid_{x=x_A} dG\left(\alpha_{i0}\mid x=x_A\right).$$

- The former is the identifiable quantity in the Blundell-Powell control function approach for cross-sectional models with endogeneity, whereas the latter is identified in the approach of Altonji and Matzkin discussed below.
- The difference between these two averages is similar to the difference between average treatment effects and average treatment effects on the treated in the program evaluation literature.

II. Integrated / weighted likelihood

- Parametric likelihood model: $f_i(\theta_0, \alpha_{i0}) = f(y_{i1}, ..., y_{iT} | x_i; \theta_0, \alpha_{i0})$, $i = 1, ..., N$.
- Interest centers in the estimation of θ or other common policy parameters.
- Central feature of this estimation problem is the presence of many nuisance parameters (the individual effects) when N is large relative to T.
- Many approaches to estimation of θ are based on an average or integrated likelihood that assigns weights to different values of α_i:

$$f_i^a(\theta) = \int f_i(\theta, \alpha_i) w_i(\alpha_i) d\alpha_i$$

  where $w_i(\alpha_i)$ is a weight, broadly defined.

- Weights may depend on θ, on the distribution of the data, as well as on covariates.
- An estimate of θ is then usually chosen to maximize the integrated likelihood of the sample under cross-sectional independence:

$$\prod_{i=1}^{N} f_i^a(\theta).$$

II.1 Fixed effects maximum likelihood

- A fixed effects approach that estimates θ jointly with the individual effects falls in this category with weights assigning all mass to $\alpha_i = \widehat{\alpha}_i(\theta)$, where $\widehat{\alpha}_i(\theta)$ is the MLE of the i-th effect for given θ.
- That is,

$$w_i(\alpha_i) = \delta(\alpha_i - \widehat{\alpha}_i(\theta))$$

  where $\delta(.)$ is Dirac's delta function.
- The resulting average likelihood in this case is just the concentrated likelihood:

$$f_i(\theta, \widehat{\alpha}_i(\theta)).$$

- In this case the weights depend on the data.

II.2 Random effects maximum likelihood

- A random effects approach is also based on an average likelihood in which the weights are chosen as a model for the distribution of individual effects in the population given covariates and initial observations.
- In this case $w_i(\alpha_i)$ is a parametric or semiparametric density or probability mass function, which does not depend on θ, but includes additional unknown coefficients:

$$w_i(\alpha_i) = g_i(\alpha_i; \xi).$$

- The integrated likelihood is the random-effects (pseudo) likelihood:

$$\int f_i(\theta, \alpha_i) g_i(\alpha_i; \xi) d\alpha_i$$

- Examples include:
  - Gaussian uncorrelated-RE ML: g is the normal density. It depends on parameters $\xi = (\mu, \sigma_\alpha^2)$.
  - Chamberlain (1984)'s correlated-effects probit: g also depends on covariates x_i.
  - Wooldridge (2005)'s approach to solving the initial conditions problem.
  - Discrete (mass point) probability distributions.

II.3 Bayesian inference

- In a Bayesian approach, an average likelihood is also constructed, choosing as weights a formulation of the prior probability distribution of α_i given θ, covariates and initial observations.
- Assuming *prior independence* conditional on θ:

$$\pi(\alpha_1, ...\alpha_N|\theta) = \pi_1(\alpha_1|\theta)...\pi_N(\alpha_N|\theta).$$

- Inference is based on the posterior:

$$\pi(\theta|y,x) \propto \pi(\theta) \prod_{i=1}^{N} \left[\int f_i\left(\theta, \alpha_i\right) \pi_i(\alpha_i|\theta) d\alpha_i \right].$$

- Weights $w_i(\alpha_i) = \pi_i(\alpha_i|\theta)$ may depend on θ and covariates.
- Random-effects specifications are a special case of hierarchical Bayesian approaches, where the prior of the effects is assumed independent of common parameters.

III. Fixed-T perspective

- All previous approaches, in general, lead to estimators of θ that are not consistent as N tends to infinity for fixed T, but have biases of order $1/T$.
- This situation, known as the "incidental parameter problem", is of particular concern when T is small relative to N, and has become one of the main challenges in modern econometrics.
- In (micro) panels typically T is much smaller than N.
- The traditional reaction to this problem has been to look for estimators yielding fixed-T consistency as N goes to infinity.
- One drawback of these methods is that they are somewhat limited to linear models and certain nonlinear models, often due to the fact that fixed-T point identification itself is problematic.
- Other considerations are that their properties may deteriorate as T increases, and that there may be superior methods that are not fixed-T consistent.

The incidental parameter problem

- The fixed effects estimator $\widehat{\theta}$ solves the first order conditions

$$\sum_{i=1}^{N} \frac{\partial \ln f_i \left(\theta, \widehat{\alpha}_i \left(\theta\right)\right)}{\partial \theta} = 0$$

  where $\widehat{\alpha}_i \left(\theta\right) = \arg\max_{\alpha} \ln f_i \left(\theta, \alpha\right)$ (based on T observations).

- Computationally ok even if N is large (the Newton-Raphson iteration decomposes nicely due to additivity of the log likelihood in the effects).

- Under standard regularity conditions $\widehat{\theta}$ is consistent if T is large:

$$\frac{1}{NT} \sum_{i=1}^{N} \frac{\partial \ln f_i \left(\theta_0, \widehat{\alpha}_i \left(\theta_0\right)\right)}{\partial \theta} \xrightarrow{p} 0 \text{ as } T \to \infty$$

  but in general

$$\operatorname*{plim}_{N \to \infty} \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial \ln f_i \left(\theta_0, \widehat{\alpha}_i \left(\theta_0\right)\right)}{\partial \theta} \neq 0.$$

- The reason is that $\widehat{\alpha}_i \left(\theta_0\right)$ is a noisy estimate of α_{i0} and the noise only goes away as T increases.

The incidental parameter problem: Example 1

- Consider $y_{it} \sim \mathcal{N}\left(\alpha_{i0}, \theta_0\right)$ so that

$$\ln f_i\left(\theta, \alpha_i\right) = k - \frac{T}{2}\ln\theta - \frac{1}{2\theta}\sum_{t=1}^{T}\left(y_{it} - \alpha_i\right)^2$$

- Here $\widehat{\alpha}_i\left(\theta\right) = \overline{y}_i$ for all θ, and

$$\widehat{\theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i\right)^2$$

- Taking a cross-sectional expectation

$$E\left(\widehat{\theta}\right) = E\left(\frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i\right)^2\right) = \theta - \frac{\theta}{T}$$

- The inconsistency only disappears as T increases.

The incidental parameter problem: Example 2

- Let $y_{it} = 1\left(\theta_0 x_{it} + \alpha_{i0} + v_{it} \geq 0\right)$ where $v_{it} \mid x_i, \alpha_{i0}$ is logistic with *cdf* $\Lambda(.)$, so that

$$\ln f_i(\theta, \alpha_i) = \sum_{t=1}^{T} \left\{ y_{it} \ln \Lambda(\theta x_{it} + \alpha_i) + (1 - y_{it}) \ln \left[1 - \Lambda(\theta x_{it} + \alpha_i)\right] \right\}$$

- Take $T = 2$ and $x_{i1} = 0, x_{i2} = 1$. Here $\widehat{\alpha}_i(\theta)$ solves the FOCs:

$$\Lambda\left(\theta x_{i1} + \widehat{\alpha}_i(\theta)\right) + \Lambda\left(\theta x_{i2} + \widehat{\alpha}_i(\theta)\right) = y_{i1} + y_{i2}.$$

- Thus, $\widehat{\alpha}_i(\theta) = \mp\infty$ if $y_{i1} + y_{i2} = 0$ or 2, and $\widehat{\alpha}_i(\theta) = -\theta/2$ if $y_{i1} + y_{i2} = 1$.
- Next, the MLE $\widehat{\theta}$ solves the FOCs from the concentrated likelihood:

$$\frac{1}{N} \sum_{i=1}^{N} 1\left(y_{i1} + y_{i2} = 1\right)\left[y_{i2} - \Lambda(\theta/2)\right] = 0,$$

  leading to

$$\widehat{\theta} = 2 \ln\left(\frac{\widehat{p}}{1 - \widehat{p}}\right),$$

  where $\widehat{p} = \widehat{\Pr}\left(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1\right) \to \Lambda(\theta_0)$ as $N \to \infty$.
- Therefore, $\widehat{\theta}$ satisfies

$$\operatorname*{plim}_{N\to\infty} \widehat{\theta} = 2\theta_0$$

- MLE estimates a relative log odds ratio that is twice as large as the truth.

Fixed effects fixed-T approaches

Fixed effects fixed-T approaches

- The general idea is separating the likelihood or at least finding a component of the likelihood that is free from the incidental parameter problem:
  - Likelihood separation: fixed-effects Poisson model.
  - Conditional likelihood: conditional logit.
- Semiparametric generalizations: Find some feature of the data (eg moments or medians) whose distribution depends on θ but not on α. These features are used to estimating θ without making assumptions about α.
  - Maximum score binary choice (Manski 1987).
  - Censored regression (Honoré 1992).
  - Dynamic binary choice (Honoré and Kyriazidou 2000).
  - Functional differencing (Bonhomme 2012).

Conditional likelihood

- Let $f_i(y_i \mid \theta, \alpha_i)$ be the likelihood for unit i. Suppose there is a statistic s_i such that

$$f_i(y_i \mid \theta, \alpha_i) \equiv f_{1i}(y_i \mid s_i, \theta, \alpha_i)\, f_{2i}(s_i \mid \theta, \alpha_i) = f_{1i}(y_i \mid s_i, \theta)\, f_{2i}(s_i \mid \theta, \alpha_i)$$

- f_{1i} is a component of the likelihood which does not depend on α_i. The idea is to base inference about θ on f_{1i} as long as there is identification.

Example 1: Linear regression

- The Gaussian linear model is

$$y_i \mid x_i, \theta_0, \alpha_i \sim \mathcal{N}\left(X_i\beta_0 + \alpha_{i0}\iota_T, \sigma_0^2 I_T\right)$$

- Letting $s_i = \overline{y}_i$, $\widetilde{y}_{it} = y_{it} - \overline{y}_i$, etc.

$$\ln f_1(y_i \mid x_i, \overline{y}_i, \theta, \alpha_i) = \ln f_1(y_i \mid x_i, \overline{y}_i, \theta) = k - \frac{(T-1)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(\widetilde{y}_{it} - \widetilde{x}_{it}\beta)^2$$

- Maximizing $\sum_{i=1}^{N}\ln f_{1i}$ wrt $\theta = (\beta, \sigma^2)$ provides WG estimates of β and bias-corrected estimates of σ^2.

Example 2: Conditional logit

- The model is
$$\Pr\left(y_{it}=1\mid x_i,\alpha_i\right)=\Lambda\left(x_{it}'\theta_0+\alpha_i\right)$$
where $\Lambda(r)=e^r/(1+e^r)$.
- Take $T=2$ to illustrate, and consider:
$$\Pr\left(y_{i1},y_{i2}\mid x_i,\alpha_i,y_{i1}+y_{i2}\right)=\begin{cases}1 & \text{if } (y_{i1},y_{i2})=(0,0)\text{ or }(1,1)\\ 1-\Lambda\left(\Delta x_{i2}'\theta_0\right) & \text{if } (y_{i1},y_{i2})=(1,0)\\ \Lambda\left(\Delta x_{i2}'\theta_0\right) & \text{if } (y_{i1},y_{i2})=(0,1)\end{cases}$$
- To see this, note that letting $z_{it}=x_{it}'\theta_0+\alpha_i$ we have
$$\Pr\left(y_{i1}=0,y_{i2}=1\mid x_i,\alpha_i,y_{i1}+y_{i2}=1\right)=\frac{\Pr\left(y_{i1}=0,y_{i2}=1\mid x_i,\alpha_i\right)}{\Pr\left(y_{i1}+y_{i2}=1\mid x_i,\alpha_i\right)}$$
$$=\frac{\left[1-\Lambda\left(z_{i1}\right)\right]\Lambda\left(z_{i2}\right)}{\left[1-\Lambda\left(z_{i1}\right)\right]\Lambda\left(z_{i2}\right)+\Lambda\left(z_{i1}\right)\left[1-\Lambda\left(z_{i2}\right)\right]}=\frac{e^{z_{i2}}}{e^{z_{i2}}+e^{z_{i1}}}=\Lambda\left(\Delta z_{i2}\right).$$
- So we obtain a binary logit likelihood for movers in which the two outcomes are $(y_{i1}=0,y_{i2}=1)$ and $(y_{i1}=1,y_{i2}=0)$ and the x's are in first differences.

Semiparametric binary choice

- Manski (1987) considered a fixed-effects binary model

$$y_{it} = 1\left(x_{it}'\theta_0 + \alpha_i + v_{it} \geq 0\right),$$

  in which the *cdf* of $-v_{it} \mid x_i, \alpha_i$ is non-parametric.

- Basic assumption:

$$\Pr(-v_{it} \leq r \mid x_i, \alpha_i) = \Pr(-v_{is} \leq r \mid x_i, \alpha_i) = F\left(r \mid x_i, \alpha_i\right) \quad \text{for all } t \text{ and } s.$$

- That is, $F\left(r \mid x_i, \alpha_i\right)$ does not change with t but is otherwise unrestricted.
- This imposes stationarity and strict exogeneity, but allows for serial dependence in v_{it}.
- Time-invariance of F implies (for $T = 2$):

$$med\left(y_{i2} - y_{i1} \mid x_i, y_{i1} + y_{i2} = 1\right) = \operatorname{sgn}\left(\Delta x_{i2}'\theta_0\right).$$

- Given $y_{i1} + y_{i2} = 1$, the difference $y_{i2} - y_{i1}$ can only equal 1 or -1. So the median will be one or the other depending on whether

$$\Pr\left(y_{i2} = 1, y_{i1} = 0 \mid x_i\right) \lesseqgtr \Pr\left(y_{i2} = 0, y_{i1} = 1 \mid x_i\right).$$

- Thus

$$\begin{aligned} med\left(\Delta y_{i2} \mid x_i, y_{i1} + y_{i2} = 1\right) &= \operatorname{sgn}[\Pr\left(y_{i2} = 1, y_{i1} = 0 \mid x_i\right) - \Pr\left(y_{i2} = 0, y_{i1} = 1 \mid x_i\right)] \\ &= \operatorname{sgn}\left[\Pr\left(y_{i2} = 1 \mid x_i\right) - \Pr\left(y_{i1} = 1 \mid x_i\right)\right]. \end{aligned}$$

- Moreover, given the model

$$\begin{aligned} \Pr\left(y_{i1} = 1 \mid x_i, \alpha_i\right) &= F\left(x_{i1}'\theta_0 + \alpha_i \mid x_i, \alpha_i\right) \\ \Pr\left(y_{i2} = 1 \mid x_i, \alpha_i\right) &= F\left(x_{i2}'\theta_0 + \alpha_i \mid x_i, \alpha_i\right), \end{aligned}$$

and monotonicity of F, we have that for any α_i (the constancy of F is crucial here):

$$\Pr\left(y_{i2} = 1 \mid x_i, \alpha_i\right) \lesseqgtr \Pr\left(y_{i1} = 1 \mid x_i, \alpha_i\right) \Leftrightarrow x_{i2}'\theta_0 \lesseqgtr x_{i1}'\theta_0.$$

- Therefore, the implication also holds on average:

$$\Pr\left(y_{i2} = 1 \mid x_i\right) \lesseqgtr \Pr\left(y_{i1} = 1 \mid x_i\right) \Leftrightarrow x_{i2}'\theta_0 \lesseqgtr x_{i1}'\theta_0.$$

Identification and estimation

- Under some conditions, θ_0 uniquely maximizes (up to scale) the expected agreement between the signs of $\Delta x'_{i2}\beta$ and Δy_{i2} conditioned on $y_{i1} + y_{i2} = 1$

$$\theta_0 = \arg\max_{\theta} E\left[\operatorname{sgn}\left(\Delta x'_{i2}\theta_0\right)\Delta y_{i2} \mid y_{i1} + y_{i2} = 1\right]$$

- Manski's identification result required an unbounded support for at least one of the explanatory variables with a non-zero coefficient.

Maximum score estimation

- This estimator selects the value that matches the signs of $\Delta x'_{i2}\theta$ and Δy_{i2} for as many observations as possible in the subsample with $y_{i1} + y_{i2} = 1$ subject to $\| \theta \| = 1$:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \operatorname{sgn}\left(\Delta x'_{i2}\theta\right)\left(y_{i2} - y_{i1}\right).$$

- The estimation criterion is unaffected by removing observations having $y_{i1} = y_{i2}$.
- It is consistent under the assumption that there is at least one unbounded continuous regressor, but it is not root-N consistent, and not asymptotically normal.

Alternative representations of the objective function

- The score objective function is

$$S_N(\theta) = \sum_{i=1}^{N} \left\{ d_{10i} 1\left(\Delta x_{i2}'\theta < 0\right) + d_{01i} 1\left(\Delta x_{i2}'\theta \geq 0\right) \right\}.$$

  where $d_{10i} = 1\left(y_{i1} = 1, y_{i2} = 0\right)$ and $d_{01i} = 1\left(y_{i1} = 0, y_{i2} = 1\right)$
- The score $S_N(\theta)$ gives the number of correct predictions we would make if we predicted $\left(y_{i1}, y_{i2}\right)$ to be $(0,1)$ whenever $\Delta x_{i2}'\theta \geq 0$.
- In contrast, $\sum_{i=1}^{N} \operatorname{sgn}\left(\Delta x_{i2}'\theta\right) \Delta y_{i2}$ gives the no. of successes minus the no. of failures.
- Median regression interpretation: minimizer of the no. of failures, which is given by

$$\frac{1}{2} \sum_{i=1}^{N} 1\left(y_{i1} \neq y_{i2}\right) \left|\Delta y_{i2} - \operatorname{sgn}\left(\Delta x_{i2}'\theta\right)\right|.$$

Smoothed Maximum Score

- Replace $S_N(\theta)$ with a smooth $S_N^*(\theta)$ whose limit *a.s.* as $N \to \infty$ is the same as $S_N(\theta)$:

$$S_N^*(\theta) = \sum_{i=1}^{N} \left\{ d_{10i} \left[1 - K\left(\Delta x_{i2}'\theta/\gamma_N\right)\right] + d_{01i} K\left(\Delta x_{i2}'\theta/\gamma_N\right) \right\}$$

  where $K(.)$ is a *cdf* and γ_N is a sequence of positive numbers with $\lim_{N\to\infty} \gamma_N = 0$.
- In this way we obtain an alternative estimator which is still not $\sqrt{N}$-consistent but is asymptotically normal (as in Horowitz, 1992).

Identification problems in binary choice with fixed T

- Useful to know which models for $\Pr(y_{i1}, ..., y_{iT} \mid x_i, \alpha_i)$ are point identified for fixed T without restricting $G(\alpha_i \mid x_i)$ and which ones are not.
- There are 2^T different possible values of $y_i = (y_{i1}, ..., y_{iT})$, denoted d_j $j = 1, ..., 2^T$. So a model is a $2^T \times 1$ vector $p(x_i, \theta, \alpha_i)$ that specifies the probabilities

$$\Pr\left(y_i = d_j \mid x_i, \theta_0, \alpha_i\right) \qquad \left(j = 1, ..., 2^T\right).$$

- Let $G_0(\alpha_i \mid x_i)$ be the true cdf. Identification will fail at θ_0 if for all x in their support, there is another cdf $G^*(\alpha_i \mid x_i)$ and $\theta^* \neq \theta_0$ in the parameter space, such that

$$\int p(x_i, \theta_0, \alpha_i)\, dG_0(\alpha_i \mid x_i) = \int p(x_i, \theta^*, \alpha_i)\, dG^*(\alpha_i \mid x_i)$$

- If so (θ_0, G_0) and (θ^*, G^*) are observationally equivalent.

- In a binary model with $\Pr(-v_{it} \leq r \mid x_i, \alpha_i) = F(r)$, if F is not logistic and x has bounded support, θ_0 suffers from local underidentification (Chamberlain 1992).
- Moreover, if x is unbounded, θ_0 is identifiable but $\sqrt{N}$-consistent estimation is possible only for the logit model.

Partial identification: set identification

- Some results for dynamic discrete choice:
  - dynamic logit: index parameters identified if $T \geq 4$.
  - dynamic probit: only set identified in general.
- In a discrete choice model where x and α are multinomial, the identified region can be written as the solution to linear programming. This is a practical way of calculating identified regions for simple models.
- Honoré and Tamer (2006) calculate identified regions in this way for an autoregressive probit model with or without a time trend or time dummies.
- The main lessons are in establishing lack of point identification for these models, and showing that, even for small values of T, the identified regions are small and tighten fast as T increases.
- Lack of identification for models with multinomial individual effects imply nonidentification of the corresponding fixed effects models.
- Set estimation and inference, a way forward (e.g. Chernozhukov, Hong, and Tamer, 2007).

Partial identification: point identification of certain marginal effects

- In a panel model, some objects of interest may be identified while others are not.

Example 1: Random coefficients model with predetermined regressor

- A simple example of identification of average effects for movers (predetermined binary regressor):

$$y_{it} = \beta_i d_{it} + \alpha_i + v_{it} \qquad E\left(v_{it} \mid d_{it}, d_{it-1}, ...\right) \qquad t = 1, 2$$

$$\begin{aligned} E\left(\Delta y_{i2} \mid d_{i1} = 0\right) &= E\left(\beta_i \mid d_{i1} = 0, d_{i2} = 1\right) \Pr\left(d_{i2} = 1 \mid d_{i1} = 0\right) \\ E\left(\Delta y_{i2} \mid d_{i1} = 1\right) &= -E\left(\beta_i \mid d_{i1} = 1, d_{i2} = 0\right) \Pr\left(d_{i2} = 0 \mid d_{i1} = 1\right) \end{aligned}$$

- $E\left(\beta_i \mid d_{i1} = 0, d_{i2} = 1\right)$ and $E\left(\beta_i \mid d_{i1} = 1, d_{i2} = 0\right)$ are identified but not $E\left(\beta_i\right)$.

Example 2: Static probit with binary regressor

- Here the common parameter θ is not point-identified. The model is

$$y_{it} = \mathbf{1}\left\{\theta x_{it} + \alpha_i \geq v_{it}\right\} \qquad v_{it} \mid x_i, \alpha_i \sim \mathcal{N}(0,1).$$

- The average effect of an increase in x_{it} from 0 to 1 is:

$$\Delta = E\left[E\left(y_{it}|x_{it}=1,\alpha_i\right) - E\left(y_{it}|x_{it}=0,\alpha_i\right)\right] = E\left[\Phi\left(\theta+\alpha_i\right) - \Phi\left(\alpha_i\right)\right].$$

- Although the overall average Δ is not point-identified for fixed T, the average effect on the subpopulation of units whose x's change over time is.
- Let us see this when $T = 2$:

$$\begin{aligned}
\Delta_{10} &= E\left[E\left(y_{i1}|x_{i1}=1,\alpha_i\right) - E\left(y_{i1}|x_{i1}=0,\alpha_i\right) \mid x_{i1}=1, x_{i2}=0\right] \\
&= E\left[E\left(y_{i1}|x_{i1}=1,x_{i2}=0,\alpha_i\right) - E\left(y_{i2}|x_{i2}=0,\alpha_i\right) \mid x_{i1}=1, x_{i2}=0\right] \\
&= E\left[E\left(y_{i1}|x_{i1}=1,x_{i2}=0,\alpha_i\right) - E\left(y_{i2}|x_{i1}=1,x_{i2}=0,\alpha_i\right) \mid x_{i1}=1, x_{i2}=0\right] \\
&= E\left[y_{i1} - y_{i2}|x_{i1}=1, x_{i2}=0\right].
\end{aligned}$$

- We have used two assumptions:
  - Strict exogeneity of x_{it}, which ensures that $E\left(y_{i1}|x_{i1}, x_{i2}, \alpha_i\right)$ and $E\left(y_{i1}|x_{i1}, \alpha_i\right)$ coincide.
  - A stationarity assumption, which implies that the conditional expectation $E\left(y_{it}|x_{it}, \alpha_i\right)$ does not depend on t (Chernozhukov, Fernandez-Val, Hahn, and Newey 2012).
- A similar result holds for the average Δ_{01} over units with $x_{i1} = 0$ and $x_{i2} = 1$.
- However, the two remaining conditional averages Δ_{00} and Δ_{11} are not point-identified.

Functional differencing

- In discrete choice models there is a large loss of information in going from the right- to the left-hand side.

- Nonlinear fixed-effects models with continuous outcomes offer greater identification opportunities (Bonhomme 2012).

- Firm-level nonlinear production functions is a relevant context of application.

- General framework: The density of $y_i = (y_{i1}, ..., y_{iT})$ conditional on x_i and α_i is given by the parametric function $f_{y_i|x_i,\alpha_i,\theta}$. The density $f_{\alpha_i|x_i}$ is left unrestricted.

Functional differencing: discrete outcomes

- Intuition: the multinomial case. Suppose that $y_i \in \{\xi_1, ..., \xi_J\}$ and $\alpha_i \in \{\zeta_1, ..., \zeta_K\}$:

$$\Pr\left(y_i = \xi_j \mid x_i\right) = \sum_{k=1}^{K} \Pr\left(y_i = \xi_j \mid x_i, \alpha_i = \zeta_k, \theta\right) \Pr\left(\alpha_i = \zeta_k \mid x_i\right)$$

- In matrix form:

$$P_{y|x} = P_x(\theta)\,\pi_x, \quad \text{for all } x,$$

  where $P_x(\theta)$ is the $J \times K$ matrix of the model probabilities for $x_i = x$, $P_{y|x}$ is the J-vector of data frequencies, and π_x the K-vector of probabilities of α_i.

- If $J \geq K$ it is easy to obtain restrictions on θ that do not involve π_x. When $P_x(\theta)$ has independent columns (for simplicity), we obtain the following restrictions on θ alone:

$$\left[I_J - P_x(\theta)\left(P_x(\theta)' P_x(\theta)\right)^{-1} P_x(\theta)'\right] P_{y|x} = 0.$$

- This "functional differencing" approach differences out the distribution of the effects.
- A differencing strategy works, even though the panel model is nonlinear, because the system that relates outcome probabilities to individual effect probabilities is linear.
- This approach delivers conditional moment restrictions for θ (given x_i) because the projection matrix above multiplies the vector of outcome probabilities.

Functional differencing: continuous outcomes

- When outcomes are continuously distributed, the matrix $P_x(\theta)$ of conditional probabilities becomes a linear mapping, or operator, which maps functions of α to functions of y.
- The image of a function $g(\alpha)$ by this operator is given by a function $L_{\theta,x}g$ of y such that:

$$[L_{\theta,x}g](y) = \int f_{y|x,\alpha}(y|x,\alpha;\theta)\, g(\alpha) d\alpha, \quad \text{for all } y.$$

- Bonhomme shows that a similar projection ("functional differencing") approach as in the discrete case can be applied in the continuous case. This approach provides conditional moment restrictions on θ that do not involve α_i.
- For these restrictions to be informative it is necessary that the image of the operator $L_{\theta,x}$ does not span the whole space of functions of y (a "non-surjective" operator).
- In the discrete case, this condition requires that the rows of the matrix $P_x(\theta)$ be linearly dependent, which is automatically satisfied provided the number of points of support of y_i exceeds that of α_i.

Random effects methods

II.2 Random effects methods

- Random effects index model:

$$y_{it} = m\left(x_{it}\theta + \alpha_i + v_{it}\right)$$

$$v_{it} \mid x_i, \alpha_i \sim \mathcal{N}\left(0, 1\right)$$

and

$$g_i\left(\alpha_i \mid x_i\right) \text{ is } \mathcal{N}\left[\lambda\left(x_i\right), \sigma_\alpha^2\right].$$

- Uncorrelated effects: $\lambda\left(x_i\right) = \mu$
- Mundlack (1978): $\lambda\left(x_i\right) = \overline{x}_i\gamma$
- Chamberlain (1984): $\lambda\left(x_i\right) = x_i'\lambda$
- Newey (1994): $\lambda\left(x_i\right)$ nonparametric.
- Altonji and Matzkin (2005): nonparametric generalization.

Mundlack's interpretation of WG

- WG can be interpreted in a much tighter random effects normal framework. In the linear model

$$y_{it} = x_{it}'\theta_0 + \alpha_i + \sigma v_{it},$$

assuming

$$v_{it} \mid x_i, \alpha_i \sim iid\mathcal{N}\left(0, 1\right)$$

and

$$\alpha_i \mid x_i \sim \mathcal{N}\left(\overline{x}_i'\gamma, \sigma_\alpha^2\right),$$

it turns out that WG maximizes

$$\int \prod_{t=1}^{T} f\left(y_{it} \mid x_i, \alpha_i\right) f\left(\alpha_i \mid x_i\right) d\alpha_i.$$

Uncorrelated random effects: linear model

- Consider a special case where there is independence between α_i and x_i ($\gamma = 0$):

$$\alpha_i \mid x_i \sim \mathcal{N}\left(0, \sigma_\alpha^2\right).$$

- In this case, letting $u_{it} = \alpha_i + \sigma v_{it}$ and $\overline{\sigma}^2 = Var\left(\overline{u}_i\right) = \sigma_\alpha^2 + \left(\sigma^2 / T\right)$, the integrated log-likelihood is

$$L\left(\beta, \sigma^2, \overline{\sigma}^2\right) = L_{WG}\left(\beta, \sigma^2\right) + L_{BG}\left(\beta, \overline{\sigma}^2\right)$$

  where

$$L_{WG}\left(\beta, \sigma^2\right) = \sum_{i=1}^{N}\left[-\frac{(T-1)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T-1}\left(y_{it}^* - x_{it}^{*\prime}\beta\right)^2\right]$$

  and

$$L_{BG}\left(\beta, \overline{\sigma}^2\right) = \sum_{i=1}^{N}\left[-\frac{1}{2}\ln\overline{\sigma}^2 - \frac{1}{2\overline{\sigma}^2}\left(\overline{y}_i - \overline{x}_i^{\prime}\beta\right)^2\right]$$

- The (uncorrelated) random effects estimator that maximizes $L\left(\beta, \sigma^2, \overline{\sigma}^2\right)$ is consistent despite correlation between x and α, but only as $T \to \infty$, because as T increases the $L_{BG}\left(\beta, \overline{\sigma}^2\right)$ component of the likelihood becomes irrelevant.
- However, when T is small it is important to allow for dependence between x and α.

Random effects probit

- The correlated random effects probit model is

$$y_{it} = 1\left(x_{it}'\theta_0 + \alpha_i + v_{it} \geq 0\right)$$

  with the same distributional assumptions as in Mundlack's model.

- However, the robustness to distributional assumptions in the linear case does not extend to binary choice.

- The uncorrelated random effects model is the special case with $\gamma = 0$.

Altonji-Matzkin's nonparametric generalization

- The model is

$$\begin{aligned} y_{it} &= m\left(x_{it}, \alpha_i, v_{it}\right) \\ \left(\alpha_i, v_{it}\right) &\perp x_i \mid \lambda\left(x_i\right) \end{aligned}$$

$g_i\left(\alpha_i \mid x_i\right) = g_i\left(\alpha_i \mid \lambda\left(x_i\right)\right)$ where $\lambda\left(x_i\right)$ is an exchangeable function of x_i (e.g. $\overline{x}_i$).

- The following average derivative effect is identified:

$$\beta\left(x_{it}\right) \equiv E_{(\alpha,v)|x_t}\left[\frac{\partial m\left(x_{it}, \alpha_i, v_{it}\right)}{\partial x_{it}} \mid x_{it}\right] = E_{\lambda|x_t}\left[\frac{\partial E\left(y_{it} \mid x_{it}, \lambda\left(x_i\right)\right)}{\partial x_{it}} \mid x_{it}\right]$$

- Note that

$$\begin{aligned} \frac{\partial E\left(y_{it} \mid x_{it}, \lambda_i\right)}{\partial x_{it}} &= \frac{\partial}{\partial x_{it}} \int_{(\alpha,v)} m\left(x_{it}, \alpha, v\right) f\left(\alpha, v \mid x_{it}, \lambda_i\right) d\left(\alpha, v\right) \\ &= \int_{(\alpha,v)} \frac{\partial m\left(x_{it}, \alpha, v\right)}{\partial x_{it}} f\left(\alpha, v \mid x_{it}, \lambda_i\right) d\left(\alpha, v\right). \end{aligned}$$

The second equality follows from the conditional exogeneity of x given λ, i.e. $\partial f\left(\alpha, v \mid x_{it}, \lambda_i\right) / \partial x_{it} = 0$.

- Exchangeability is a strong assumption.
- Basic idea is conditioning on $\lambda\left(x_i\right)$ as a substitute for conditioning on α_i.

Dynamic discrete choice panel models

Dynamic discrete choice panel models
Introduction

- Prototypical model is

$$y_{it} = 1\left(\alpha y_{i(t-1)} + \beta x_{it} + \eta_i + v_{it} \geq 0\right)$$

$$v_{it} \mid x_i, \eta_i, y_{i(t-1)}, ..., y_{i1} \sim iid\ F$$

- This is a model for

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

- The lagged dependent variable $y_{i(t-1)}$ captures "state dependence" and is "fixed-effects endogenous" by construction.
- The external regressor x_{it} is also fixed-effects endogenous but strictly exogenous with respect to v_{it}.

Spurious state dependence

- Unobserved heterogeneity may cause spurious state dependence. That is, we might have no genuine state dependence:

$$\Pr\left(y_{it} = 1 \mid y_{i(t-1)}, \eta_i\right) = \Pr\left(y_{it} = 1 \mid \eta_i\right)$$

but spurious state dependence

$$\Pr\left(y_{it} = 1 \mid y_{i(t-1)}\right) \neq \Pr\left(y_{it} = 1\right)$$

just because

$$\Pr\left(y_{it} = 1 \mid y_{i(t-1)}\right) = \int \Pr\left(y_{it} = 1 \mid \eta_i\right) dG\left(\eta_i \mid y_{i(t-1)}\right)$$

and η_i depends on $y_{i(t-1)}$.

Dynamic discrete choice panel models vs. duration models

- The previous model can be regarded as a convenient discrete duration model for exits from two states:

$$\begin{aligned} h_u(x,\eta) &= \Pr\left(y_{it}=1 \mid y_{i(t-1)}=0, x_i, \eta_i\right) = F\left(\beta x_{it}+\eta_i\right) \\ h_e(x,\eta) &= \Pr\left(y_{it}=0 \mid y_{i(t-1)}=1, x_i, \eta_i\right) = 1-F\left(\alpha+\beta x_{it}+\eta_i\right) \end{aligned}$$

  where $h_u(x,\eta)$ is the exit rate from state 0 into state 1 (e.g. exit rate from unemployment) while $h_e(x,\eta)$ is the exit rate from state 1 into state 0 (e.g. exit rate from employment).

- Note that

$$\frac{\partial h_u(x,\eta)}{\partial x_j} \Big/ \frac{\partial h_u(x,\eta)}{\partial x_k} = \frac{\beta_j}{\beta_k} = -\frac{\partial h_e(x,\eta)}{\partial x_j} \Big/ \frac{\partial h_e(x,\eta)}{\partial x_k}$$

  So, as a model for durations the specification above has the unattractive property that relative effects from the two exit rates are equal but with opposite signs.

- An example of a more flexible specification in this context is in Card and Hyslop (2005) discussed below.

The initial conditions problem in dynamic models

- We have $f(y_1, ..., y_T \mid x, \eta)$. To do random effects we integrate:

$$f(y_1, ..., y_T \mid x) = \int f(y_1, ..., y_T \mid x, \eta)\, dG(\eta \mid x)$$

- Now consider

$$f(y_1, ..., y_T \mid x, \eta) = \prod_{t=2}^{T} f(y_t \mid y_{t-1}, x, \eta)\, f(y_1 \mid x, \eta).$$

- If we proceed as above the density $f(y_1 \mid x, \eta)$ needs to be specified, which may not be available. This is the so called "initial conditions problem".
- Typically, we just have specified a model for the transitions $f(y_t \mid y_{t-1}, x, \eta)$.
- $f(y_1 \mid x, \eta)$ could be chosen as the steady state distribution. One problem is that the steady state may be unknown or may not exist. Another problem is that we may not wish to impose stationarity in estimation even if available.
- Alternatively, we could start from

$$f(y_2, ..., y_T \mid y_1, x, \eta) = \prod_{t=2}^{T} f(y_t \mid y_{t-1}, x, \eta)$$

  and integrate using $G(\eta \mid y_1, x)$:

$$f(y_2, ..., y_T \mid y_1, x) = \int \prod_{t=2}^{T} f(y_t \mid y_{t-1}, x, \eta)\, dG(\eta \mid y_1, x).$$

- Doing this save us having to specify $f(y_1 \mid x, \eta)$ but requires us to specify $G(\eta \mid y_1, x)$ as opposed to $G(\eta \mid x)$.

Fixed T consistent dynamic models

- Conditional logit does not work with lagged dependent variables or other predetermined variables. It requires independence of all x's on the transitory errors, but there is still a fixed T fixed effects approach available under certain circumstances.

Autoregressive logit (Chamberlain)

- The model is

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

- Consider $T = 4$. The main result is

$$\Pr\left(y_{i2} = 1 \mid y_{i4}, y_{i2} + y_{i3} = 1, y_{i1}, \eta_i\right) = \Lambda\left[\alpha\left(y_{i1} - y_{i4}\right)\right],$$

  which does not depend on η.

- Therefore, sequences of the form $(y_1, 0, 0, y_4)$ or $(y_1, 1, 1, y_4)$ drop out of the conditional likelihood.

- Contributions of the form $(y_1, 1, 0, y_4)$ and $(y_1, 0, 1, y_4)$ are retained in principle. But of those, observations with $y_1 = y_4$ are not informative about α.

- We are allowed to only retain $(y_1 = 1, y_4 = 0)$ and $(y_1 = 0, y_4 = 1)$ because we are conditioning on these random variables.

- So we end up with 4 different types of informative contributions:

$$
\begin{aligned}
(1,1,0,0) &\longrightarrow \quad \frac{e^{\alpha}}{1+e^{\alpha}} = p, \text{ say} \\
(0,1,0,1) &\longrightarrow \quad \frac{e^{-\alpha}}{1+e^{-\alpha}} = \frac{1}{1+e^{\alpha}} = 1-p \\
(1,0,1,0) &\longrightarrow \quad \frac{1}{1+e^{\alpha}} = 1-p \\
(0,0,1,1) &\longrightarrow \quad \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}} = p
\end{aligned}
$$

- Let $n_1 = \#(1,1,0,0)$, $n_2 = \#(0,1,0,1)$, $n_3 = \#(1,0,1,0)$, $n_4 = \#(0,0,1,1)$, and let the total number of usable observations be $n_5 = n_1 + n_2 + n_3 + n_4$.
- So we can estimate p as

$$\widehat{p} = \frac{n_1 + n_4}{n_5}$$

and

$$\widehat{\alpha} = \ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = \ln\left(\frac{n_1+n_4}{n_2+n_3}\right)$$

- Population wise we have $\alpha = \ln(p_A/p_B)$ where

$$
\begin{aligned}
p_A &= \Pr\{(1,1,0,0) \text{ or } (0,0,1,1)\} \\
p_B &= \Pr\{(0,1,0,1) \text{ or } (1,0,1,0)\}.
\end{aligned}
$$

Honoré and Kyriazidou's method

- Their basic model is

$$\Pr\left(y_{it} = 1 \mid y_i^{t-1}, x_i, \eta_i\right) = \Lambda\left(\alpha y_{i(t-1)} + \beta x_{it} + \eta_i\right)$$

- The following is the central result:

$$\Pr\left(y_{i2} = 1 \mid y_{i4}, y_{i2} + y_{i3} = 1, y_{i1}, x_i, x_{i3} = x_{i4}, \eta_i\right) = \Lambda\left[\alpha\left(y_{i1} - y_{i4}\right) + \beta\left(x_{i2} - x_{i3}\right)\right]$$

- The method conditions on $\Delta x_{i4} = 0$ in addition to the autoregressive-logit type of conditioning.
- Identification relies on variation in Δx_{i3} and in lack of variation in Δx_{i4}.
- If x is discrete root-N consistent estimation is possible, but not if x is continuous.
- We may think of this estimation problem as based on two functions of $x_{i1}, \Delta x_{i2}, \Delta x_{i3}, \Delta x_{i4}$ (or just $\Delta x_{i3}, \Delta x_{i4}$) for $(y_1 = 1, y_4 = 0)$ and $(y_1 = 0, y_4 = 1)$.
- Effectively, estimation of the model's parameters is based on a nonparametric estimate of a conditional expectation at one particular value $(\Delta x_{i4} = 0)$.

Random effects approaches for discrete choice dynamic models

- These include:
  - Models with autocorrelation that are estimated by simulation (Hajivassiliou and Ruud, 1994).
  - Extensions of Chamberlain (1984)'s method to observed lagged dependent variables, latent lagged dependent variables and general predetermined variables.
- Latent lagged dependent variables: Arellano, Bover, and Labeaga (1999) for censored VAR models.
- Binary choice with general predetermined variables: Arellano and Carrasco (2003).
- Observed lagged dependent variables: Wooldridge (2005)
  - The idea is to specify the density of the effects given strictly exogenous x's and initial conditions.

Illustration: Effect of a time-limited earnings subsidy on welfare participation (Card and Hyslop, 2005)

The SSP experiment (Self Sufficiency Project, 1992–1995, Canada)
The following program was designed:

- Out of concern that the welfare system was promoting long-term dependency.
- The target group was single parents that were welfare recipients for at least one year.
- Those selected for the policy, become "eligible" for subsidy payments if they manage to get a full time job within a year of selection.
- Once they are eligible, they can move back and forth between work and welfare. When they are at (full time) work, they are entitled to subsidy payments, for 3 years from the time of the first payment. After that, they return to regular welfare conditions.
- The subsidy is substantial. Some monthly figures are:
  - maximum welfare grant: \$712
  - minimum wage job for 30 hours per week: \$650
  - min. wage + SSP subsidy $= 650 + \frac{1}{2}(2500 - 650) = \$1575$
  - gain from welfare to work without SSP = –\$62
  - gain with SSP = \$863
- SSP used a randomized design in two different locations:
  - Control group: 2826 single parents (95.3% women, aged 32 on average)
  - Program group: 2858 (of which 34% were eventually eligible for subsidies)

Results of the experiment

- Figures 1a and 3 in the paper summarize the situation:
  - Figure 3 shows employment rates of controls and treatments for the duration of the program (approx 4 years): very large employment effects around the time of eligibility, followed by declining effects until a full collapse and the end of the program.
  - Figure 1a tells a similar story for welfare participation rates.
- The SSP experiment produced one of the largest impacts on welfare participation ever recorded in the experimental evaluation literature. At peak, SSP produced a 14 percentage point reduction in welfare participation.
- The bad news is that SSP had no permanent impact, giving no support to the idea that temporary wage subsidies can have a permanent effect on program dependency (presumably through the development of work habits, labor market experience, etc.).

A baseline empirical model for welfare participation for controls

- Let $y = 1$ if person is a welfare participant. Card and Hyslop say they "adopt *a panel data approach* rather than a *hazard modelling approach* because of the high incidence of multiple spells in our data".

$$\Pr(y_{it} = 1 \mid y_{it-1}, y_{it-2}, x_{it}, \alpha_i) =$$
$$\Lambda(x_{it}\beta + (\gamma_{10} + \gamma_{11}\alpha_i) y_{it-1} + (\gamma_{20} + \gamma_{21}\alpha_i) y_{it-2} + (\gamma_{30} + \gamma_{31}\alpha_i) y_{it-1}y_{it-2} + \alpha_i)$$
$$(t = 1, ..., T = 69)$$

$$P\left(y_{i1}, ..., y_{iT} \mid y_{i0}, y_{i(-1)}, x_i, \alpha_i\right) = \prod_{t=1}^{T} P(y_{it} \mid y_{it-1}, y_{it-2}, x_{it}, \alpha_i)$$

$$P\left(y_{i1}, ..., y_{iT} \mid y_{i0}, y_{i(-1)}, x_i\right) = \int P\left(y_{i1}, ..., y_{iT} \mid y_{i0}, y_{i(-1)}, x_i, \alpha_i\right) dF\left(\alpha_i \mid y_{i0}, y_{i(-1)}, x_i\right)$$

- The only x is time since random assignment (a fourth order polynomial)
- Because of the design, everyone has $y_{i0} = y_{i(-1)} = 1$. Thus, $F\left(\alpha_i \mid y_{i0}, y_{i(-1)}, x_i\right)$ does not vary with $y_{i0}, y_{i(-1)}, x_i$ and we write just $F(\alpha_i)$ for shortness.
- If $\gamma_{k1} = 0$, for $k = 1, 2, 3$ the degree of state dependence is restricted to be invariant to the unobserved heterogeneity.
- Almost half of the sample have just one spell on welfare. For many individuals in the sample the ML estimate of α_i is $+\infty$.

This model specifies the following different transitions:

- Transition or exit rate (from work to welfare) in the first month of a work spell:

$$\Lambda\left(x_{it}\beta + (\gamma_{20} + \gamma_{21}\alpha_i) + \alpha_i\right)$$

- Transition rate (from work to welfare) in subsequent months of a work spell:

$$\Lambda\left(x_{it}\beta + \alpha_i\right)$$

- Transition rate (from welfare to work) in the first month of a welfare spell:

$$1 - \Lambda\left(x_{it}\beta + (\gamma_{10} + \gamma_{11}\alpha_i) + \alpha_i\right)$$

- Transition rate (from welfare to work) in subsequent months of a welfare spell:

$$1 - \Lambda\left(x_{it}\beta + (\gamma_{10} + \gamma_{11}\alpha_i) + (\gamma_{20} + \gamma_{21}\alpha_i) + (\gamma_{30} + \gamma_{31}\alpha_i) + \alpha_i\right)$$

They do a detailed and informative goodness of fit analysis.

Joint model of welfare participation and eligibility for SSP payments for treatments

- The model for treatments is

$$\begin{aligned} & \Pr\left(y_{it}=1 \mid y_{it-1}, y_{it-2}, x_{it}, \alpha_i, E_{it}, t_i^e\right) \\ = \quad & \Lambda\left[x_{it}\beta+\left(\gamma_{10}+\gamma_{11}\alpha_i\right) y_{it-1}+\left(\gamma_{20}+\gamma_{21}\alpha_i\right) y_{it-2}\right. \\ & \left.+\left(\gamma_{30}+\gamma_{31}\alpha_i\right) y_{it-1} y_{it-2}+\alpha_i+\tau_{it}\right] \end{aligned}$$

where

$$\tau_{it}=\tau\left(t, E_{it}, t_i^e, y_{it-1}\right)$$

and $E_{it}=1$ if eligible at the beginning of month t.

A model of the eligibility process that accounts for the potential correlation between the probability of entering or leaving welfare and the probability of attaining SSP eligibility.

- This is a hazard model for the event of establishing eligibility in month t, conditional on not establishing it earlier:

$$\Pr\left(E_{it} \mid E_{it-1}, E_{it-2}, \ldots, x_{it}, \alpha_i\right)=\begin{cases}\Phi\left[d(t)-k\left(\alpha_i\right)\right] & \text{if } E_{it-1}=0 \text{ and } t \leq 14 \\ 1 & \text{if } E_{it-1}=1 \\ 0 & \text{if } E_{it-1}=0 \text{ and } t>14\end{cases}$$

- Therefore, the model recognizes that E_{it} is an endogenous explanatory variable in the sense that it is correlated with α_i. We have

$$
\begin{aligned}
& P\left(y_{i1}, ..., y_{iT}, E_{i1}, ..., E_{iT} \mid y_{i0}, y_{i(-1)}, x_i, \alpha_i\right) \\
= & \prod_{t=1}^{T} P\left(y_{it}, E_{it} \mid y_{it-1}, y_{it-2}, E_{it-1}, x_{it}, \alpha_i\right)
\end{aligned}
$$

$$
= \prod_{t=1}^{T} P\left(y_{it} \mid y_{it-1}, y_{it-2}, E_{it}, x_{it}, \alpha_i\right) \Pr\left(E_{it} \mid E_{it-1}, E_{it-2}, ..., x_{it}, \alpha_i\right)
$$

and

$$
\begin{aligned}
& P\left(y_{i1}, ..., y_{iT}, E_{i1}, ..., E_{iT} \mid y_{i0}, y_{i(-1)}, x_i\right) \\
= & \int \prod_{t=1}^{T} P\left(y_{it} \mid y_{it-1}, y_{it-2}, E_{it}, x_{it}, \alpha_i\right) \Pr\left(E_{it} \mid E_{it-1}, E_{it-2}, ..., x_{it}, \alpha_i\right) dF\left(\alpha_i\right).
\end{aligned}
$$

Experimental versus nonexperimental effects

- The point of the paper (similar to Ham and LaLonde, 1996) is that, although the experimental comparisons between the treatment and control groups remain valid, the interpretation of such impacts is confounded by the different treatment effects associated with two different sets of incentives:
  - An *entitlement effect* that makes you lower your reservation wage (and hence increase your exit rate from welfare) while you still have a chance of attaining the eligibility status.
  - An *establishment effect* for those enjoying eligibility status that leads to a lower reservation wage relative to controls and the non-established treated.
- These effects are clear from a theoretical model of the welfare-work decision that serves to guide the formulation of the empirical model.
- Treatment status is independent of α_i by construction, but treatment status is not independent of α_i conditionally on $E_{it} = 1$. Thus, $F(\alpha_i)$ is the same for treatments and controls but $F(\alpha_i \mid E_{it})$ is not.
- Card and Hyslop claim that although their model is not structural (utility based), it can be used to evaluate the impacts of alternative subsidy programs.

Bayesian methods

Bayesian methods

Integration versus simulation

- A classical approach to estimation is to maximize the log-average likelihood wrt (θ, ξ), which requires computing integrals with respect to α.
- In nonlinear panels the integrals are generally not available in closed form and must be approximated numerically (using quadrature or simulation-based approaches).
- The Bayesian connection suggests another way to estimate θ. Indeed, random-effects ML coincides with the posterior mode of θ, where the prior for α_i is $g_i(\alpha_i; \xi)$, and (θ, ξ) have independent flat (improper) priors.
- So, an alternative approach is to generate a Markov chain of parameter draws using these priors, which may be interpreted as a computationally convenient way of calculating random-effects ML estimates.
- The statistical equivalence between Bayesian and classical approaches is not limited to posterior mode with flat priors. Any non-dogmatic priors on (θ, ξ) will result in large-N asymptotically equivalent estimates.
- Using posterior mean instead of posterior mode has asymptotically negligible effects.
- Advances in computation have made Bayesian methods increasingly attractive from an applied perspective. Leading to a pragmatic Bayesian-frequentist synthesis, as MCMC methods are viewed as a way of computing estimators with a frequentist justification.
- Bayesian techniques are also useful for computing frequentist confidence intervals.

Markov Chain Monte Carlo (MCMC) methods applied to panel models

- MCMC methods are used to generate a (recursive) sequence of draws from the posterior distribution of the model's parameters, starting with initial parameter values.

- The posterior corresponds to the equilibrium distribution of the Markov chain, which is reached after a sufficiently large number of steps.

- The output of the chain is interpreted as a sequence of draws from the parameters' posterior distribution, so that its features (mean, mode..) can be directly computed.

- In a panel context, it is often convenient to treat $\alpha_1, ..., \alpha_N$ as additional parameters that are drawn jointly with (θ, ξ). The s-th step of the chain may take the form:

  - Update $\xi^{(s)}$ given $\alpha_1^{(s-1)}, ..., \alpha_N^{(s-1)}$. This step treats the draws of individual effects obtained in the previous step as observations.
  - For each $i = 1, ..., N$, update $\alpha_i^{(s)}$ given y_i, x_i, $\theta^{(s-1)}$, and $\xi^{(s)}$.
  - Update $\theta^{(s)}$ given $y_1, ..., y_N$, $x_1, ..., x_N$, and $\alpha_1^{(s)}, ..., \alpha_N^{(s)}$. To draw θ, the researcher proceeds as if the individual effects were observed.

- Metropolis-Hastings methods are typically used here.

- An appealing feature is that the output of the Markov chain does not only provide estimates of θ and ξ, but also asymptotically valid frequentist confidence intervals.

Average marginal effects

- Common parameters aside, we are interested in averages of individual quantities taken over the distribution of (x_i, α_i). The general form for some known function $m()$ is:

$$M = E_{(x_i,\alpha_i)}\left[m\left(x_i, \alpha_i; \theta\right)\right].$$

- Examples are moments of the distribution of individual effects: $m_i(\theta, \alpha_i) = \alpha_i^k$, or the marginal effect of a covariate in a probit model: $m_i(\theta, \alpha_i) = \theta_k \frac{1}{T}\sum_{t=1}^{T} \phi\left(x_{it}'\theta + \alpha_i\right)$.
- A first approach to estimate M is to replace in the expectation the distribution of individual effects by its random-effects estimate. This results in the following estimate:

$$\widehat{M} = \frac{1}{N}\sum_{i=1}^{N}\int m\left(x_i, \widehat{\alpha}_i; \widehat{\theta}\right) g_i\left(\alpha_i; \widehat{\xi}\right) d\alpha_i.$$

- Under correct specification, $\widehat{M}$ is root-N consistent. Numerical integration is required.
- An alternative estimate may be computed from the outcome of a Markov chain. MCMC will deliver a sequence of draws of θ and $\alpha_1, .., \alpha_N$, from which it is easy to get a sequence of draws from the posterior distribution of the average marginal effect

$$M_N\left(\theta, \alpha_1, ..., \alpha_N\right) = \frac{1}{N}\sum_{i=1}^{N} m\left(x_i, \alpha_i; \theta\right).$$

- A natural estimate is then the posterior mode, or mean, of $M_N\left(\theta, \alpha_1, ..., \alpha_N\right)$.
- When $g_i\left(\alpha_i; \xi\right)$ is misspecified, the posterior mean (or mode) of M_N is large-T consistent while $\widehat{M}$ is not. This is due to the impact of the prior of α_i on the posterior of M tending to disappear as $T \to \infty$.

Bias-reduction methods

IV. An alternative population framework: non-fixed T perspective

- Often T is much smaller than N and this situation has justified the mainstream approach, which treats data as a multivariate sample from a cross-sectional population with a fixed number of observations per unit.

- However, there are also panels in which T may not be negligible from the point of view of time series inference, and not negligible relative to N, even if N may still be much larger than T. For example, N may be small relative to T^3.

- An alternative approach in those situations is to think of the data as a realization from a random field in which neither T nor N are fixed.

- This is an alternative population framework where statistical learning from individual time series is not ruled out, so it may lead to different conclusions on what quantities are identified.

Non-fixed T asymptotic properties

- Let $\widehat{\theta}$ be a fixed effects estimator that maximizes some concentrated (pseudo) log likelihood $\sum_{i=1}^{N}\sum_{t=1}^{T}\ln f_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)$ and let $\theta_T = \text{plim}_{N\to\infty}\widehat{\theta}$.
- In general $\theta_T \neq \theta_0$, but usually for smooth objective functions

$$\theta_T = \theta_0 + \frac{B}{T} + O\left(\frac{1}{T^2}\right).$$

- Under standard regularity conditions $\widehat{\theta} - \theta_T$ is asymptotically normal as $N, T \to \infty$:

$$\sqrt{NT}\left(\widehat{\theta} - \theta_T\right) \xrightarrow{d} \mathcal{N}\left(0, V\right)$$

  where V is the large-T asymptotic variance of $\widehat{\theta}$.
- Under these conditions $\widehat{\theta}$ is centered at θ_0 if $N/T \to 0$ but it is asymptotically biased if T grows at the same rate as N. If $N/T \to c > 0$ and $N/T^3 \to 0$:

$$\sqrt{NT}\left(\widehat{\theta} - \theta_0 - \frac{B}{T}\right) \xrightarrow{d} \mathcal{N}\left(0, V\right).$$

- Thus, unless $N/T \approx 0$, asymptotic confidence intervals based on $\widehat{\theta}$ will be incorrect, due to the limiting distribution of $\sqrt{NT}\left(\widehat{\theta} - \theta_0\right)$ not being centered at 0.

Bias-reduced estimation

- The aim in this literature has been to obtain estimators of θ with biases of order $1/T^2$ (as opposed to $1/T$) and similar large-sample dispersion as the corresponding uncorrected methods when T/N tends to a constant. That is, find $\widetilde{\theta}$ that satisfies

$$\widetilde{\theta} = \widehat{\theta} - \frac{B}{T} + o_p(1).$$

- This is done in the hope that the reduction in the order of magnitude of the bias will essentially eliminate the incidental parameter problem, even in panels where T is much smaller than N.
- An interesting property of panel data estimators is that bias reduction happens with no increase in the asymptotic variance as N/T tends to a constant.
- To obtain sufficiently accurate confidence intervals from this type of asymptotic approximation, the bias should be small relative to the standard deviation.
    - For first-order bias corrected estimators, this requires that N be small relative to T^3 (e.g. N small relative to $1,000$ or to $8,000$ for $T = 10$ or 20, respectively).

- Similar to the bias of the fixed effects estimand $\theta_T - \theta_0$, the bias in the expected fixed effects score at θ_0 can be expanded in orders of magnitude of T:

$$E\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\ln f_{it}\left(\theta_0,\widehat{\alpha}_i\left(\theta_0\right)\right)\right]=\frac{b_i\left(\theta_0\right)}{T}+o\left(\frac{1}{T}\right)$$

  and also the bias in the expected concentrated likelihood in a neighborhood of θ_0:

$$E\left[\frac{1}{T}\sum_{t=1}^{T}\ln f_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{T}\sum_{t=1}^{T}\ln f_{it}\left(\theta,\overline{\alpha}_i\left(\theta\right)\right)\right]=\frac{\beta_i\left(\theta\right)}{T}+o\left(\frac{1}{T}\right)$$

  where $\overline{\alpha}_i\left(\theta\right)=\mathrm{plim}_{T\to\infty}\widehat{\alpha}_i\left(\theta\right)$ uniformly in θ.

- These expansions motivate alternative approaches to bias correction based on adjusting
  - the estimator (Hahn and Newey 2004, Hahn and Kuersteiner 2011),
  - the estimating equation (Woutersen 2002, Arellano 2003, Carro 2007),
  - or the objective function (Arellano and Hahn 2007, Bester and Hansen 2009).

- Each of them based on analytical or simulation-based approximations to the bias.

Bias-reducing priors

- A different approach to bias reduction is in Arellano and Bonhomme (2009). They consider estimators that maximize an integrated likelihood

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln \int f_i(\theta, \alpha_i) \, w_i(\alpha_i) \, d\alpha_i$$

and describe the class of weights $w_i(\alpha_i)$ that produce first-order unbiased estimators.

- The idea is to look for priors such that the corresponding estimator has $B = 0$.
- It turns out that bias reducing priors depend on the data in general, unless an orthogonal reparameterization is available.
- Bayesian techniques can be used for estimation.
- Asymptotically valid (as $N, T \to \infty$) confidence intervals can be read from the posterior distribution of θ.

Random effects

- In general RE ML is not bias reducing. Exceptions are:
  - a) The true population distribution of the effects belongs to the postulated family.
  - b) Gaussian RE ML is bias reducing in models that are linear in the individual effects.
  - c) Individual effects and common parameters are information orthogonal.
- The RE ML bias depends on the Kullback-Leibler distance between the population distribution of the effects and its best approximation in the random effects family.

Automatic bias reduction: jackknife approaches

- In addition to analytical approaches and weighted likelihood approaches, the literature has emphasized automatic approaches to bias reduction.
- In static panel models, Hahn and Newey (2004) propose the *delete-one jackknife*:

$$\widetilde{\theta} = T\widehat{\theta} - (T-1)\frac{1}{T}\sum_{t=1}^{T}\widehat{\theta}_{(t)}$$

  or

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widetilde{B}}{T}$$

  where $\widehat{\theta}_{(t)}$ is the FE estimator based on the subsample excluding the t-th period observation, and

$$\frac{\widetilde{B}}{T} = (T-1)\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\theta}_{(t)} - \widehat{\theta}\right)$$

- To see why this works consider

$$\theta_T = \theta_0 + \frac{B}{T} + \frac{D}{T^2} + O\left(\frac{1}{T^3}\right)$$

$$\theta_{T-1} = \theta_0 + \frac{B}{T-1} + \frac{D}{(T-1)^2} + O\left(\frac{1}{(T-1)^3}\right)$$

$$T\theta_T - (T-1)\theta_{T-1} = \theta_0 + \left(\frac{1}{T} - \frac{1}{T-1}\right)D + O\left(\frac{1}{T^2}\right) = \theta_0 + O\left(\frac{1}{T^2}\right)$$

Jackknife approaches (continued)

- Hahn and Newey (2004) proved that $\sqrt{NT}\left(\widetilde{\theta}-\theta_0\right)$ has the same asymptotic variance as $\sqrt{NT}\left(\widehat{\theta}-\theta_0\right)$ when $N/T \to c$ and no asymptotic bias.
- The *delete-last-observation* approach is not to be recommended as it will remove bias but increase variance (ie using $\widehat{\theta}_{(T)}$ as the sample analog for θ_{T-1}).

Dynamic models

- The *split-panel jackknife* method of Dhaene and Jochmans (2006) allows for dynamics and predetermined variables.
- The idea is to obtain the fixed-effects estimator on the two subsamples $[1, T/2]$ and $[T/2+1, T]$ (assuming T even for simplicity).
- Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ denote the two estimates, and let $\widehat{\theta}$ denote the estimate based on the full sample. The first-order bias term of $\widehat{\theta}_1$ is $B/(T/2) = 2B/T$, while that of $\widehat{\theta}$ is B/T. Thus, the following estimator is unbiased to first order:

$$\widehat{\theta}^R = 2\widehat{\theta} - \frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2}.$$

- Split-panel jackknife estimators have the same asymptotic variance as the MLE and no asymptotic bias when $N/T \to c$.
- Dhaene and Jochmans also show that within the class of split-panel jackknife estimators, the half-panel jackknife estimator $\widehat{\theta}^R$ minimizes all higher-order bias terms.

Jackknife approaches (continued)

- Jackknife bias-corrected estimates of average marginal effects can be readily obtained.
- Split-panel jackknife relies on stationarity and this rules out aggregate time effects.
- Fernández-Val and Weidner (2011) discuss a generalized jackknife approach to deal simultaneously with individual and time effects.

Finite sample performance of bias-reduction estimators

- The available evidence on the finite-sample performance of the various approaches to bias reduction is encouraging.
- In static and dynamic settings that mimic PSID data (e.g. Carro 2007), these techniques tend to remove at least half of the bias, while keeping the variance virtually unchanged.
- An issue concerns the possibility to reduce the bias further. Second-order bias reduction can be simply implemented using a variant of the split-panel jackknife approach. However, the Monte Carlo evidence presented in Dhaene and Jochmans suggests that higher-order bias reduction may be associated with increased variance.
- There is so far too little comparison of the various bias reduction approaches on simulated data.
- Moreover, although panel data bias reduction has been used in some empirical applications (e.g. Fernández-Val and Vella 2011, Hospido 2012), more applications are needed.

Concluding remarks

- The random effects perspective is a *general* estimation approach.
- Link between classical RE and Bayesian approaches. Worth stressing because MCMC methods are convenient for computing RE estimates and their confidence intervals.
- RE approaches, however, rely on parametric assumptions on the distribution of REs. When violated, RE estimates are subject to an incidental parameter problem, just as fixed-effects MLE. As a result, RE estimators are generally fixed T inconsistent.
- Point identification may fail when T is fixed and the distribution of REs is left unrestricted. In discrete choice panel models, parameters are typically set-identified.
- However, in models with continuous outcomes, panel data offer opportunities for point-identification that remain largely unexplored.
- When T is not negligible relative to N, it makes sense to view incidental parameter problems as TS finite-sample bias. In general, RE estimates are consistent as $T \to \infty$.

Concluding remarks (continued)

- The first-order bias of RE MLE is a function of the (Kullback-Leibler) distance between the true RE density and its best approximation in the parametric family.
- This characterization suggests that one may achieve bias reduction by letting the parametric distribution of REs become increasingly flexible as $N \to \infty$.
- In the absence of covariates, this is within reach but in the presence of covariates, however, achieving the required level of "flexibility" so as to remove the first-order bias on the parameter of interest is more challenging.