On the identification and estimation of panel data dynamic discrete choice models

Bo E. Honoré

Princeton University

ESAM, July 2009
Based on published, unpublished and unwritten papers with

- Ekaterini Kyriazidou
- Elie Tamer
- Anna D’Addio
- Adriana Lleras–Muney
Old Question (Heckman)

Panel Data Discrete Choice Model with Heterogeneity and State Dependence:

\[ y_{it} = 1 \{ x_{it} \beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geq 0 \} \quad t = 1, \ldots, T; \quad i = 1, \ldots, n \]

Issue: With short panel, can we distinguish the effect of \( \gamma y_{i,t-1} \) from the effect of \( \alpha_i \).

Some answers out there, but still a very active research area.

The aim is to add to this literature.
Motivation (1)

Simple models for relationship between health and labor force participation

\[ y_{it} = 1 \left\{ x_{it-1} \beta_2 + y_{it-1} \gamma_2 + \delta_i + \nu_{it} \geq 0 \right\} \]
\[ x_{it} = 1 \left\{ x_{it-1} \beta_1 + y_{it-1} \gamma_1 + \alpha_i + \varepsilon_{it} \geq 0 \right\} \]

or

\[ y_{it} = 1 \left\{ x_{it} \beta_2 + y_{it-1} \gamma_2 + \delta_i + \nu_{it} \geq 0 \right\} \]
\[ x_{it} = 1 \left\{ x_{it-1} \beta_1 + y_{it-1} \gamma_1 + \alpha_i + \varepsilon_{it} \geq 0 \right\} \]

Would probably want to include age-trends or dummies. And perhaps other explanatory variables.
Motivation

Notation gets out of hand, so consider the simpler model

\[ z_{it} = 1 \{ z_{it-1} \gamma_1 + z_{it-2} \gamma_2 + \delta_i + u_{it} \geq 0 \} \]

This is a special case of

\[ y_{it} = 1 \{ x_{it} \beta_2 + y_{it-1} \gamma_2 + \delta_i + \nu_{it} \geq 0 \} \]
\[ x_{it} = 1 \{ x_{it-1} \beta_1 + y_{it-1} \gamma_1 + \alpha_i + \varepsilon_{it} \geq 0 \} \]

with \( \beta_2 = \gamma_1 \), \( \beta_1 = \gamma_2 \) and \( \delta_i = \alpha_i \).

To see this let

\[ z_{i,2t} = y_{it} \quad \text{and} \quad z_{i,2t-1} = x_{it} \]
\[ u_{i,2t} = \nu_{it} \quad \text{and} \quad u_{i,2t-1} = \varepsilon_{it} \]
Also of independent interest

Simple model for labor force participation

\[ y_{it} = 1 \left\{ y_{it-1} \gamma_i + y_{it-2} \gamma_{y_{it-1}} + \delta_i + \nu_{it} \right\} \geq 0 \]

Note that \( \gamma_0 = 0 \) implies

\[ P \left( y_{it} = 1 \mid y_{it-1} = 0, y_{it-2} \right) = \Lambda \left( \delta_i \right) \]

(transition probability unrestricted except that it does not depend on state at time \( t - 2 \): no duration dependence)

Might want to include other explanatory variables.
The Issues

In general (but not always):

- $n$ large, $T$ small: asymptotics with $n \to \infty$ for $T$ fixed.
- We cannot just difference out the “fixed” effects $(\delta_i, \alpha_i)$.
- Estimating $(\delta_i, \alpha_i)$ typically does not work ($T$ fixed).
- A “random effects” approach also gets us into problems.

This talk

- Highlight the issues and show some solutions.
- Then get back to motivating examples.
A bit of background

Consider the (seemingly) simpler model

\[ y_{it} = 1 \{ x'_{it} \beta + y_{i,t-1} \gamma + \alpha_i + \epsilon_{it} \geq 0 \} \]

where \( \{x_{it}\} \) is strictly exogenous.

Interested in estimation of \((\gamma, \beta)\).

More generally

\[ f \left( y_{it} \mid y_{it-1}, \ldots, y_{it-k}, x_i, \alpha_i \right) \]

where \( x_i = \{x_{it}\}_{t=1}^{T} \).
Parametric approach ("random effects")

Specify $f(\alpha_i)$ and $f(y_{i1}|x_i, \alpha_i)$

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

But what is $f(y_{i1}|x_i, \alpha_i)$?

With stationarity and time–invariant $x_i$, one can sometimes solve for it. But these are strong assumptions.
Practical solutions:

- **Heckman’s solution**: Specify a flexible functional form for $f(y_{i1}|x_i, \alpha_i)$

- **Wooldridge’s solution**: Specify $f(\alpha_i|x_i, y_{i1})$.
  - if done right, it can fit into a standard Stata command

- **The disadvantage**: With unbalanced panel we end up specifying $f(\alpha_i|x_{i1}, ..., x_{iT}, y_{i1})$ and $f(\alpha_i|x_{i1}, ..., x_{iT+1}, y_{i1})$. So we are making implicit assumptions on how $\{x_{it}\}$ evolves.

*Alternative*: Fixed effects approach.
Less parametric approach ("fixed effects"): 

\[ y_{it} = 1 \{ x_{it}' \beta + y_{i,t-1} \gamma + \alpha_i + \varepsilon_{it} \geq 0 \} \]

In a linear world, we would first-difference and get 

\[ y_{it} = (x_{it} - x_{it-1})' \beta + (y_{i,t-1} - y_{i,t-2}) \gamma + (\varepsilon_{it} - \varepsilon_{it-1}) \geq 0 \]

and do IV.

But differencing would not eliminate \( \alpha_i \); and we don’t really know how to do IV in nonlinear models
Conditioning on a sufficient statistic

Logit with fixed effects ($T = 2$):

$$P(y_{i1} = d_1, y_{i2} = d_2 | x_{i1}, x_{i2}, \alpha_i)$$

$$= \frac{\exp(x_{i1}\beta + \alpha_i)^{d_1}}{1 + \exp(x_{i1}\beta + \alpha_i)} \cdot \frac{\exp(x_{i2}\beta + \alpha_i)^{d_2}}{1 + \exp(x_{i2}\beta + \alpha_i)}$$

By simple calculation

$$P(y_{i1} = d | x_{i1}, x_{i2}, \alpha_i, y_{i1} + y_{i2} = 1)$$

$$= \frac{\exp(((x_{i1} - x_{i2})\beta)^d}{1 + \exp(((x_{i1} - x_{i2})\beta)}$$

(Rasch, 1960)
This is a special case of a more general idea

Suppose

- the distribution of $y_i$ conditional on $(S_i, x_i, \alpha_i)$ does not depend on $\alpha_i$.
- the distribution of $y_i$ conditional on $(S_i, x_i, \alpha_i)$ depends on $\theta$,

then one can estimate $\theta$ by maximum likelihood using the conditional distribution of the data given $(S_i, x_i)$.

(Andersen, 1970).
AR(1) Logit

\[ y_{it} = 1 \{ y_{it-1} \gamma + \delta_i + \nu_{it} \geq 0 \}, \quad t = 2, 3, \ldots, T \geq 4 \]

Let

\[ s_1 = \sum_{t=1}^{T} y_t \]

Cox (old paper) and Chamberlain (1985) showed that conditional on \((y_1, s_1, y_T)\), the distribution of \((y_{i1}, \ldots, y_{iT})\) does not depend on \(\delta_i\), but it does depend on \(\gamma\).
**AR(2) logit**

Model from before

\[ y_{it} = 1 \{ y_{it-1} \gamma_{1i} + y_{it-2} \gamma_2 + \delta_i + \nu_{it} \geq 0 \} \quad t = 3, \ldots, T \geq 6 \]

Let

\[ s_1 = \sum_{t=1}^{T} y_t \quad \text{and} \quad s_{11} = \sum_{t=2}^{T} y_t y_{t-1} \]

Chamberlain (1984) showed that conditional on \((y_1, y_2, s_1, s_{11}, y_{T-1}, y_T)\), the distribution of \((y_{i1}, \ldots, y_{iT})\) does not depend on \((\delta_i, \gamma_{1i})\), but it does depend on \(\gamma_2\).

Questions:

- But what if we wanted to estimate a common \(\gamma_1\)?
- And what if the errors were normal?
AR(2)–more–general–logit

\[ y_{it} = 1 \left\{ y_{it-1} \gamma_{1i} + y_{it-2} \gamma_{2,y_{i,t-1}} + \delta_i + \nu_{it} \geq 0 \right\} \quad t = 3, \ldots, T \geq 6 \]

As before let

\[ s_1 = \sum_{t=1}^{T} y_t \quad \text{and} \quad s_{11} = \sum_{t=2}^{T} y_t y_{t-1} \]

Turns out that conditional on \((y_1, y_2, s_1, s_{11}, y_{T-1}, y_T)\), the distribution of \((y_{i1}, \ldots, y_{iT})\) does not depend on \((\delta_i, \gamma_{1i})\), but it does depend on \((\gamma_{2,0}, \gamma_{2,1})\).
Even more general model

Now consider the case where $\gamma_{2,1}$ is also individual–specific. In that case we would need to also condition on

$$s_{111} = \sum_{t=3}^{T} y_t y_{t-1} y_{t-2}$$

which would give us an expression that depends on $\gamma_{2,0}$ but not on the individual–specific effects (D’Addio and Honoré, under revision).

Link to duration literature...
Back to model of interest

The same kind of stuff works for the models in the paper with Lleras–Muney:

\[ y_{it} = \begin{cases} 1 \{ x_{it} \beta_2 + y_{it-1} \gamma_2 + \delta_i + \nu_{it} \geq 0 \} \\ x_{it} = \begin{cases} 1 \{ x_{it-1} \beta_1 + y_{it-1} \gamma_1 + \alpha_i + \epsilon_{it} \geq 0 \} \end{cases} \]

- With logistic errors, we can identify \( \gamma_2 \) and \( \beta_1 \) while allowing \( \gamma_1 \) and \( \beta_2 \) to be individual–specific.
- But not clear how to identify \( \gamma_1 \) and \( \beta_2 \).
But what about covariates?

For example if $\varepsilon_{it}$ is i.i.d. logistic, then

$$P(y_{i2} = 1 | y_{i1}, y_{i2} + y_{i3} = 1, y_{i4}, x_{i3} = x_{i4}) = \frac{\exp((x_{i2} - x_{i3})\beta + \gamma(y_{i1} - y_{i4}))}{1 + \exp((x_{i2} - x_{i3})\beta + \gamma(y_{i1} - y_{i4}))}$$

which does not depend on $\alpha_i$.

Idea: Match on $x_{i2} \approx x_{i3}$. A bit like matching.

This idea can be applied to the models considered here.

BUT: Need the distribution of $x_{i2} - x_{i3}$ to have support in a neighborhood of 0. Time-dummies/trends are ruled out.
All very confusing

- All of this require some kind of "trick"
- It would be nice to think about these models from a more general point of view.
Back to basics

\[ y_{it} = 1 \{ x_{it}' \beta + y_{i,t-1} \gamma + \alpha_i + \varepsilon_{it} \geq 0 \} \]

Let \( p_0(\alpha, x^T) = P(y_{i0} = 1 \mid x_i^T, \alpha_i) \) and let \( \theta \) be all the parameters of the model (incl. parameters in distribution of \( \varepsilon_{it} \) and \( \alpha_i \)).
Identified Region

The set of \((p_0 (\cdot, \cdot), \theta)\) that are consistent with the data-generating process, is

\[
\{ (p_0 (\cdot, \cdot), \theta) : P \left( \pi \left( A; p_0 (\cdot, x^T), \theta \right) = P \left( A| x^T \right) \right) = 1 \text{ for all } A \}
\]

and the sharp bounds on \(\theta\) is given by

\[
\{ \theta : \exists p_0 (\cdot, \cdot) \text{ such that } P \left( \pi \left( A; p_0 (\cdot, x^T), \theta \right) = P \left( A| x^T \right) \right) = 1 \text{ for all } A \}
\]
The identified region is the solution to a number of optimization problems. For example

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

where $A$ is the set of all outcomes.
Or

\[
\max_{p_0(\cdot, \cdot), \theta} E \left[ w \left( x^T \right) \log \left( \pi \left( y_i; p_0 \left( \cdot, x^T \right), x^T, \theta \right) \right) \right] =
\]

\[
\max_{p_0(\cdot, \cdot), \theta} E \left[ w \left( x^T \right) \log \left( \int p_0 \left( \alpha, x^T \right)^{y_{i1}} \left( 1 - p_0 \left( \alpha, x^T \right) \right)^{1-y_{i1}} \prod_{t=2}^{T} P \left( y_{it} | x_i^T, y_{it-1}; \theta \right) dG \left( \alpha | x_i^T; \theta \right) \right) \right]
\]

Essentially what people do (old Heckman papers).
Useful computational approach (if time permits)

Suppose that $\alpha$ has a discrete distribution with known points of support, $a_m$, and unknown probabilities $\rho_m$. Ignore $x_i^T$. Then
\[ \pi (A; p_0 (\cdot), \theta) = \sum_{m=1}^{M} \rho_m (p_0 (a_m) \pi (A | y_0 = 1, a_m; \theta) + (1 - p_0 (a_m)) \pi (A | y_0 = 0, a_m; \theta)) \]

\[ = \sum_{m=1}^{M} z_{m,1} \pi (A | y_0 = 1, a_m; \theta) + \sum_{m=1}^{M} z_{m,0} \pi (A | y_0 = 0, a_m; \theta) \]

where

\[ z_{m,1} = \rho_m p_0 (a_m) \quad \text{and} \quad z_{m,0} = \rho_m (1 - p_0 (a_m)) \]

\(\{z_m\} \) gives probabilities in the joint distribution of \(y_0\) and \(\alpha\)
Θ is the values of θ for which the equations

\[ \sum_{m=1}^{M} \sum_{\ell=0}^{1} z_{m,\ell} \pi ( A | y_0 = \ell, a_m; \theta) = P(A) \]  \quad (1) 

\[ \sum_{m=1}^{M} \sum_{\ell=0}^{1} z_{m,\ell} = 1 \]  \quad (2)

\[ z_{m,\ell} \geq 0 \]  \quad (3)

have a solution for \( \{ z_m \}_{m=1}^{2M} \).
\[ \Theta = \arg \max_{\theta} \maximize_{\{z_m\}, \{v_i\}} \sum_i -v_i \]

\[ P(A) = \sum_{m=1}^{M} \sum_{\ell=0}^{1} z_{m,\ell} \pi (A | y_0 = \ell, a_m; \theta) = v_A \]

\[ 1 - \sum_{m=1}^{M} \sum_{\ell=0}^{1} z_{m,\ell} = v_0 \]
\[ z_{m,\ell} \geq 0 \]
\[ v_i \geq 0 \]

(The optimal function value is 0).
Example:

\[ y_{it} = \begin{cases} 1 & \{ y_{i,t-1} \gamma + t \beta + \alpha_i + \varepsilon_{it} - 0.35 \geq 0 \} \quad \text{for} \quad t = 1, 2, \ldots, T \end{cases} \]

with \( \varepsilon_{it} \) i.i.d. standard normal. \( \alpha_i \) discretized normal.

Note that time–trend or time–dummies will be relevant for application to health outcome.
\( \gamma = 0.50, \, \beta = 0.10, \, T = 3 \)
Marginal Effect: \((0.0965, 0.1489)\)

\( \gamma = 0.50, \, \beta = 0.10, \, T = 4 \)
Marginal Effect: \((0.1262, 0.1265)\)
Recall:

We considered

\[ y_{it} = 1 \{ y_{it-1} \gamma_{1i} + y_{it-2} \gamma_2 + \delta_i + \nu_{it} \geq 0 \} \]

where \( \nu_{it} \) is i.i.d. logistic.

The distribution of \( (y_{i1}, \ldots, y_{iT}) \) conditional on

\[ \left( y_1, y_2, \sum_{t=1}^{T} y_t, \sum_{t=2}^{T} y_t y_{t-1}, y_{T-1}, y_T \right) \]

does not depend on \( (\delta_i, \gamma_{1i}) \), but it does depend on \( \gamma_2 \). So we can estimate \( \gamma_2 \).

But we did not know about \( \gamma_1 \) (assuming that it is the same across observations).

*Numerical calculations suggest that \( \gamma_1 \) is point-identified.*
We did not know about

$$y_{it} = 1 \left\{ y_{it-1} \gamma_{1i} + y_{it-2} \gamma_{2i} + \delta_i + \nu_{it} \geq 0 \right\}$$

when $\nu_{it}$ is normal. *Calculated the identified region for a specific case*
$$\gamma_{21} = 0.30, \quad \gamma_{22} = 0.20$$
Conclusions

- **Bad News**
  - The whole literature is confusing
  - The model we care about not point identified.

- **Good news**
  - We can do things
  - Non-identification may not be so bad (if we knew how to do the inference)