Binary Choice Model for Panel data

José Fajardo
FGV/EBAPE

Panel Data Binary Choice Models

Random Utility Model for Binary Choice

$U_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}$ + Person i specific effect

Fixed effects using "dummy" variables

$U_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}$

Random effects using omitted heterogeneity

$U_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i$

Same outcome mechanism: $Y_{it} = 1[U_{it} > 0]$

Ignoring Unobserved Heterogeneity

Assuming strict exogeneity; $\text{Cov}(\mathbf{x}_{it}, u_i + \varepsilon_{it}) = 0$

$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$

$\text{Prob}[y_{it} = 1 \mid x_{it}] = \text{Prob}[u_i + \varepsilon_{it} > -\mathbf{x}_{it}'\boldsymbol{\beta}]$

Using the same model format:

$\text{Prob}[y_{it} = 1 \mid x_{it}] = F\left(\mathbf{x}_{it}'\boldsymbol{\beta} / \sqrt{1+\sigma_u^2}\right) = F(\mathbf{x}_{it}'\boldsymbol{\delta})$

This is the 'population averaged model.'

Ignoring Heterogeneity (Broadly)

- Presence will generally make parameter estimates look smaller than they would otherwise.
- Ignoring heterogeneity will definitely distort standard errors.
- Partial effects based on the parametric model may not be affected very much.
- Is the pooled estimator 'robust?' Less so than in the linear model case.

Panel Probit Model

The "Panel Probit Model"

The German innovation data: T=5 (N=1270)

$y_{it}^* = x_{it}'\boldsymbol{\beta} + \varepsilon_{it}$, $y_{it} = 1[x_{it}'\boldsymbol{\beta} + \varepsilon_{it} > 0]$

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1T} \\ \rho_{12} & 1 & \dots & \rho_{2,T} \\ \dots & \dots & \dots & \dots \\ \rho_{1T} & \rho_{2,T} & \dots & 1 \end{pmatrix} \right]$$

FIML

$$\log L = \sum_{i=1}^{N} \log \text{Prob}[y_{i1},\ldots,y_{i5}]$$
$$= \sum_{i=1}^{N} \log \int_{-\infty}^{(2y_{i5}-1)\mathbf{x}_{i5}'\boldsymbol{\beta}} \cdots \int_{-\infty}^{(2y_{i1}-1)\mathbf{x}_{i1}'\boldsymbol{\beta}} g(\mathbf{v} \mid \boldsymbol{\Sigma}^*) dv_1 \ldots dv_5$$
$$g(\mathbf{v} \mid \boldsymbol{\Sigma}^*) = (2\pi)^{-5/2} \mid \boldsymbol{\Sigma}^* \mid^{-1/2} \exp[-(1/2)\mathbf{v}'(\boldsymbol{\Sigma}^*)^{-1}\mathbf{v}]$$
$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 1 & q_{i1}q_{i2}\rho_{12} & \cdots & q_{i1}q_{i5}\rho_{15} \\ q_{i1}q_{i2}\rho_{12} & 1 & \cdots & q_{i2}q_{i2}\rho_{22} \\ \cdots & \cdots & \cdots & \cdots \\ q_{i1}q_{i5}\rho_{1T} & q_{i2}q_{i5}\rho_{2,T-1} & \cdots & 1 \end{pmatrix}$$
$$q_{it} = 2y_{it} - 1$$

See Greene, W., "Convenient Estimators for the Panel Probit Model: Further Results," Empirical Economics, 29, 1, Jan. 2004, pp. 21-48.

GMM

From the marginal distributions:

$E[y_{it} - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta}) \mid \mathbf{X}_i] = 0$ (note: strict exogeneity)

Suggests orthogonality conditions

$$E\begin{bmatrix} (y_{i1} - \Phi(\mathbf{x}_{i1}'\boldsymbol{\beta}))\mathbf{x}_{i1} \\ (y_{i2} - \Phi(\mathbf{x}_{i2}'\boldsymbol{\beta}))\mathbf{x}_{i2} \\ \cdots \\ (y_{i5} - \Phi(\mathbf{x}_{i5}'\boldsymbol{\beta}))\mathbf{x}_{i5} \end{bmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{pmatrix}$$ 5*K moments.

Application: Health Care Panel Data

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods

Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. **There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).**

Variables in the file are

→	DOCTOR	= 1(Number of doctor visits > 0)
	HOSPITAL	= 1(Number of hospital visits > 0)
	HSAT	= health satisfaction, coded 0 (low) - 10 (high)
	DOCVIS	= number of doctor visits in last three months
	HOSPVIS	= number of hospital visits in last calendar year
→	PUBLIC	= insured in public health insurance = 1; otherwise = 0
	ADDON	= insured by add-on insurance = 1; otherswise = 0
	HHNINC	= household nominal monthly net income in German marks / 10000. (4 observations with income=0 were dropped)
	HHKIDS	= children under age 16 in the household = 1; otherwise = 0
	EDUC	= years of schooling
	AGE	= age in years
	MARRIED	= marital status

Pooled vs. RE Panel Estimator

```
Binomial Probit Model
Dependent variable               DOCTOR
--------+--------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+--------------------------------------------------------------
Constant|    .02159          .05307         .407   .6842
     AGE|    .01532***       .00071       21.695   .0000      43.5257
    EDUC|   -.02793***       .00348       -8.023   .0000      11.3206
  HHNINC|   -.10204**        .04544       -2.246   .0247       .35208
--------+--------------------------------------------------------------
Unbalanced panel has   7293 individuals
--------+--------------------------------------------------------------
Constant|   -.11819          .09280       -1.273   .2028
     AGE|    .02232***       .00123       18.145   .0000      43.5257
    EDUC|   -.03307***       .00627       -5.276   .0000      11.3206
  HHNINC|    .00660          .06587         .100   .9202       .35208
     Rho|    .44990***       .01020       44.101   .0000
--------+--------------------------------------------------------------
```

Example10.do

Partial Effects

```
-----------------------------------------------------------------------
Partial derivatives of E[y] = F[*]  with
respect to the vector of characteristics
They are computed at the means of the Xs
Observations used for means are All Obs.
--------+--------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]  Elasticity
--------+--------------------------------------------------------------
        |Pooled
     AGE|    .00578***       .00027       21.720   .0000       .39801
    EDUC|   -.01053***       .00131       -8.024   .0000      -.18870
  HHNINC|   -.03847**        .01713       -2.246   .0247      -.02144
--------+-----------------------------------------------------------
        |Based on the panel data estimator
     AGE|    .00620***       .00034       18.375   .0000       .42181
    EDUC|   -.00918***       .00174       -5.282   .0000      -.16256
  HHNINC|    .00183          .01829         .100   .9202       .00101
--------+--------------------------------------------------------------
```

Effect of Clustering

- Y_{it} must be correlated with Y_{is} across periods
- Pooled estimator ignores correlation
- Broadly, $y_{it} = E[y_{it}|\mathbf{x}_{it}] + w_{it}$,
  - $E[y_{it}|\mathbf{x}_{it}] = \text{Prob}(y_{it} = 1|\mathbf{x}_{it})$
  - w_{it} is correlated across periods
- Assuming the marginal probability is the same, the pooled estimator is consistent. (We just saw that it might not be.)
- Ignoring the correlation across periods generally leads to underestimating standard errors.

Cluster Correction: Doctor

```
---------------------------------------------------------------------
Binomial Probit Model
Dependent variable                DOCTOR
Log likelihood function     -17457.21899
--------+------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+------------------------------------------------------------
        | Conventional Standard Errors
Constant|     -.25597***        .05481      -4.670   .0000
     AGE|      .01469***        .00071      20.686   .0000      43.5257
    EDUC|     -.01523***        .00355      -4.289   .0000      11.3206
  HHNINC|     -.10914**         .04569      -2.389   .0169       .35208
  FEMALE|      .35209***        .01598      22.027   .0000       .47877
--------+------------------------------------------------------------
        | Corrected Standard Errors
Constant|     -.25597***        .07744      -3.305   .0009
     AGE|      .01469***        .00098      15.065   .0000      43.5257
    EDUC|     -.01523***        .00504      -3.023   .0025      11.3206
  HHNINC|     -.10914*          .05645      -1.933   .0532       .35208
  FEMALE|      .35209***        .02290      15.372   .0000       .47877
--------+------------------------------------------------------------
```

Panel Logit Model

Conditional Estimation

- Principle: $f(y_{i1},y_{i2},\ldots \mid \text{some statistic})$ is free of the fixed effects for some models.
- Maximize the conditional log likelihood, given the statistic.
- Can estimate $\boldsymbol{\beta}$ without having to estimate α_i.
- Only feasible for the logit model. (Poisson and a few other continuous variable models. No other discrete choice models.)

Binary Logit Conditional Probabilities

$$\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it})=\frac{e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

$$\text{Prob}\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2},\ldots,Y_{iT_i}=y_{iT_i}\,\middle|\,\sum_{t=1}^{T_i}y_{it}\right)$$

$$=\frac{\exp\left(\sum_{t=1}^{T_i}y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it}=S_i}\exp\left(\sum_{t=1}^{T_i}d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}=\frac{\exp\left(\sum_{t=1}^{T_i}y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\substack{\text{All }\binom{T_i}{S_i}\text{ different ways that}\\ \Sigma_t d_{it}\text{ can equal }S_i}}\exp\left(\sum_{t=1}^{T_i}d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

Denominator is summed over all the different combinations of T_i values of y_{it} that sum to the same sum as the observed $\Sigma_{t=1}^{T_i}y_{it}$. If S_i is this sum, there are $\binom{T}{S_i}$ terms. May be a huge number. An algorithm by Krailo and Pike makes it simple.

Example: Two Period Binary Logit

$$\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it})=\frac{e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

$$\text{Prob}\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2},\ldots,Y_{iT_i}=y_{iT_i}\,\middle|\,\sum_{t=1}^{T_i}y_{it},\text{data}\right)=\frac{\exp\left(\sum_{t=1}^{T_i}y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it}=S_i}\exp\left(\sum_{t=1}^{T_i}d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

$$\text{Prob}\left(Y_{i1}=0,Y_{i2}=0\,\middle|\,\sum_{t=1}^{2}y_{it}=0,\text{data}\right)=1.$$

$$\text{Prob}\left(Y_{i1}=1,Y_{i2}=0\,\middle|\,\sum_{t=1}^{2}y_{it}=1,\text{data}\right)=\frac{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})+\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}$$

$$\text{Prob}\left(Y_{i1}=0,Y_{i2}=1\,\middle|\,\sum_{t=1}^{2}y_{it}=1,\text{data}\right)=\frac{\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})+\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}$$

$$\text{Prob}\left(Y_{i1}=1,Y_{i2}=1\,\middle|\,\sum_{t=1}^{2}y_{it}=2,\text{data}\right)=1.$$

Fixed Effects Logit Health Model: Conditional vs. Unconditional

Table 2.14 Estimated Fixed Effects Logit Models

	Unconditional Estimator				Conditional Estimator				
	LogL = -8506.164				LogL = -5669.541				
	LogLR = -17365.15								
	7293 Individuals								
	3289 Individuals Bypassed								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
AGE	-.1095	.0076	-14.405	.0000	-.0881	.0068	-12.984	.0000	43.5257
EDUC	.0090	.0835	.108	.9141	.0126	.0718	.176	.8604	11.3206
INCOME	.6038	.1968	3.068	.0022	.4767	.1750	2.724	.0064	.35208
MARRIED	-.1091	.1114	-.979	.3276	-.0772	.0983	-.785	.4322	.75862
KIDS	-.0167	.0793	-.210	.8337	-.0059	.0706	-.084	.9331	.40273
	Partial Effects				Partial Effects				
AGE	-.0288	.0063	-4.102	.0000	-.0012	.00009	-13.961	.0000	43.5257
EDUC	.0021	.0193	.110	.9122	.0002	.0010	.176	.8605	11.3206
INCOME	.1429	.0582	2.455	.0141	.0066	.0023	2.920	.0035	.35208
MARRIED	-.0258	.0015	-17.531	.0000	-.0011	.0014	-.789	.4303	.75862
KIDS	-.0038	.0008	-5.225	.0000	-.00008	.0010	-.084	.9331	.40273

Example10.do

Estimating Partial Effects

"The fixed effects logit estimator of $\boldsymbol{\beta}$ immediately gives us the effect of each element of $\mathbf{x}_i$ on the **log-odds ratio**… Unfortunately, we cannot estimate the partial effects… unless we plug in a value for α_i. Because the distribution of α_i is unrestricted – in particular, $E[\alpha_i]$ is not necessarily zero – **it is hard to know what to plug in for α_i.** In addition, we cannot estimate average partial effects, as doing so would require finding $E[\Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i)]$, a task that apparently requires specifying a distribution for α_i."
(Wooldridge, 2002)

Example10.do

Advantages and Disadvantages of the FE Model

- Advantages
  - Allows correlation of effect and regressors
  - Fairly straightforward to estimate
  - Simple to interpret
- Disadvantages
  - Model may not contain time invariant variables
  - Not necessarily simple to estimate if very large samples (Stata just creates the thousands of dummy variables)
  - **The incidental parameters problem: Small T bias**

Incidental Parameters Problems: Conventional Wisdom

- General: The unconditional MLE is biased in samples with fixed *T* except in special cases such as linear or Poisson regression (even when the FEM is the right model).
  The conditional estimator (that bypasses estimation of α_i) is consistent.
- Specific: Upward bias (experience with probit and logit) in estimators of $\boldsymbol{\beta}$

Bias Correction Estimators

- Motivation: Undo the incidental parameters bias in the fixed effects probit model:
  - (1) Maximize a penalized log likelihood function, or
  - (2) Directly correct the estimator of β
- Advantages
  - For (1) estimates α_i so enables partial effects
  - Estimator is consistent under some circumstances
  - (Possibly) corrects in dynamic models
- Disadvantage
  - No time invariant variables in the model
  - Practical implementation
  - Extension to other models? (Ordered probit model (maybe) – see JBES 2009)

Bias Reduction

- Parametric (probit and logit) models with fixed effects
- Recent references: All about probit and logit models.
  - [1] Carro, J., "Estimating dynamic panel data discrete choice models with fixed effects," JE, 140, 2007, pp. 503-528
  - [2] Val, F., "Fixed Effects estimation of structural parameters and marginal effects in panel probit models," JE, 2010
  - [3] Hahn, J. and G. Kuersteiner, "Bias reduction for dynamic nonlinear panel models with fixed effects," UCLA, 2003
  - [4] Hahn, J. and W. Newey, "Jackknife and Analytical Bias reduction for nonlinear panel models," Econometrica, 2004.
  - See, also, bibliographies and work of T. Woutersen, B. Honoré and E. Kyriazidou.

A Mundlak Correction for the FE Model

Fixed Effects Model :

$y_{it}^{*} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}$, i = 1,...,N; t = 1,...,T_i

$y_{it} = 1$ if $y_{it} > 0$, 0 otherwise.

Mundlak (Wooldridge, Heckman, Chamberlain), ...

$\alpha_i = \gamma + \boldsymbol{\theta}'\bar{\mathbf{x}}_i + u_i$ (Projection, not necessarily conditional mean)

where u is normally distributed with mean zero and standard deviation σ_u and is uncorrelated with $\bar{\mathbf{x}}_i$ or $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$

Reduced form random effects model

$y_{it}^{*} = \gamma + \boldsymbol{\theta}'\bar{\mathbf{x}}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i$, i = 1,...,N; t = 1,...,T_i

$y_{it} = 1$ if $y_{it} > 0$, 0 otherwise.

Mundlak Correction

Table 2.17 Random Effects Model with Mundlak Correction

	Random Effects Probit				Group Means Addition				
	LogL = -15424.40				LogL = -15404.26				
	LogL0 = -17365.76				LogL0 = -17365.76				
	7293 Individuals								
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.9459	.1116	8.473	.0000	.6551	.1232	5.320	.0000	43.5257
AGE	-.0365	.0015	-24.279	.0000	-.0521	.0036	-14.582	.0000	43.5257
EDUC	.0817	.0073	11.230	.0000	.0031	.0421	.073	.9415	11.3206
INCOME	.3207	.0717	4.474	.0000	.2937	.0959	3.064	.0022	.35208
MARRIED	.0188	.0346	.544	.5863	-.0429	.0534	-.803	.4220	.75862
KIDS	.0430	.0298	1.443	.1490	-.0019	.0397	-.048	.9614	.40273
AGEBAR					.0193	.0039	4.895	.0000	
EDUCBAR					.0790	.0427	1.848	.0646	
INCMBAR					.3451	.1696	2.037	.0211	
MARRBAR					.0499	.0717	.695	.4871	
KIDSBAR					.0936	.0616	1.520	.1285	
Rho	.5404	.0100	53.842	.0000	.5389	.0100	53.822	.0000	

Fixed Effects Models Summary

- Incidental parameters problem if T < 10 (roughly)
- Inconvenience of computation
- Appealing specification
- Alternative semiparametric estimators?
  - Theory not well developed for T > 2
  - Not informative for anything but slopes (e.g., predictions and marginal effects)
- Ignoring the heterogeneity definitely produces an inconsistent estimator (*even with cluster correction*!)
- Mundlak correction is a useful common approach.

Escaping the FE Assumptions

Chamberlain (again)

Structure

$\text{Prob}(y_{it} = 1 \mid x_{it}) = F(\alpha_i + x_{it}'\beta),\ \alpha_i = \alpha + \bar{\mathbf{x}}_i'\delta + w_i$

Reduced form is a random effects model

$\text{Prob}(y_{it} = 1 \mid x_{it}) = F(\alpha + \bar{\mathbf{x}}_i'\delta + x_{it}'\beta + w_i)$

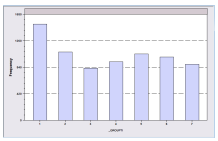
(Does not allow time invariant effects (again).)

Estimation:

(1) FIML

(2) Period by period, then reconcile with minimum distance

Unbalanced Panels



Group Sizes

Most theoretical results are for balanced panels.

Most real world panels are unbalanced.

Often the gaps are caused by attrition.

The major question is whether the gaps are 'missing completely at random.' If not, the observation mechanism is endogenous, and at least some methods will produce questionable results.

Researchers rarely have any reason to treat the data as nonrandomly sampled. (This is good news.)

Unbalanced Panels and Attrition 'Bias'

- Test for 'attrition bias.' (Verbeek and Nijman, Testing for Selectivity Bias in Panel Data Models, International Economic Review, 1992, 33, 681-703.
  - Variable addition test using covariates of presence in the panel
  - Nonconstructive – what to do next?
- Do something about attrition bias. (Wooldridge, Inverse Probability Weighted M-Estimators for Sample Stratification and Attrition, Portuguese Economic Journal, 2002, 1: 117-139)
  - Stringent assumptions about the process
  - Model based on probability of being present in each wave of the panel

Panel Data and Selection

Selection equation with time invariant individual effect

$d_{it} = 1[\mathbf{z}_{it}'\boldsymbol{\gamma} + \theta_i + \eta_{it} > 0]$

Observation mechanism: $(y_{it}, \mathbf{x}_{it})$ observed when $d_{it} = 1$

Primary equation of interest

Common effects linear regression model

$y_{it} \mid (d_{it} = 1) = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$

"Selectivity" as usual arises as a problem when the unobservables are correlated; $\text{Corr}(\varepsilon_{it}, \eta_{it}) \neq 0$.

The common effects, θ_i and α_i make matters worse.

Panel Data Sample Selection Models

Verbeek, Economics Letters, 1990.

$d_{it} = 1[\mathbf{z}_{it}'\boldsymbol{\gamma} + w_i + \eta_{it} > 0]$ **(Random effects probit)**

$y_{it} \mid (d_{it} = 1) = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$; **(Fixed effects regression)**

Proposed "marginal likelihood" based on joint normality

$$\log L_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{t=1}^{T_i} \Phi\left[(2d_{it}-1)\frac{\mathbf{z}_{it}'\boldsymbol{\gamma} + \Delta_{it} + u_{i,1} + d_{it}u_{i,2}}{\sqrt{\sigma_\eta^2(1-d_{it}\rho^2)}}\right] f(u_{i,1}, u_{i,2})du_{i,1}du_{i,2}$$

$\Delta_{it} = (\rho/\sigma_\varepsilon)d_{it}\left[(y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta}\right]$

(Integrate out the random effects; difference out the fixed effects.)

$u_{i,1}, u_{i,2}$ are time invariant uncorrelated standard normal variables

How to do the integration? Natural candidate for simulation.

(Not mentioned in the paper. Too early.)

[Verbeek and Nijman: Selectivity "test" based on this model, International Economic Review, 1992.]

Selection with Fixed Effects

$$y_{it}^* = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},\ \eta_i = \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \tau w_i, w_i \sim N[0,1]$$

$$d_{it}^* = \theta_i + \mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it},\ \theta_i = \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i, v_i \sim N[0,1]$$

$$(\varepsilon_{it}, u_{it}) \sim N_2[(0,0),(\sigma^2, 1, \rho\sigma)].$$

$$L_i = \int_{-\infty}^{\infty}\prod_{d_{it}=0}\Phi\left[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i'\boldsymbol{\delta} - \omega v_i\right]\phi(v_i)dv_i$$

$$\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{d_{it}=1}\Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it}}{\sigma}\right)\phi_2(v_i, w_i)dv_i dw_i$$

$$\varepsilon_{it} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \tau w_i$$

Attrition

- In a panel, t=1,…,T individual I leaves the sample at time K_i and does not return.
- If the determinants of attrition (especially the unobservables) are correlated with the variables in the equation of interest, then the now familiar problem of sample selection arises.

Dealing with Attrition

- The attrition issue: Appearance for the second interview was low for people with initial low QOL (death or depression) or with initial high QOL (don't need the treatment). Thus, missing data at exit were clearly related to values of the dependent variable.
- Solutions to the attrition problem
  - Heckman selection model (used in the study)
    - Prob[Present at exit|covariates] = $\Phi(z'\theta)$ (Probit model)
    - Additional variable added to difference model $\lambda_i = \Phi(\mathbf{z}_i'\theta)/\Phi(\mathbf{z}_i'\theta)$
  - The FDA solution: fill with zeros. (!)

Methods of Estimating the Attrition Model

- Heckman style "selection" model
- Full information maximum likelihood
- Weighting schemes that account for the "survivor bias"

Dynamic Models

A Dynamic RE Probit

Habit persistence and latent heterogeneity

$y_{i,t} = 1[\mathbf{x}_{i,t}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{i,t} > 0]$, t = 1,...,T; i=1,...,N

Fixed effects assumption; $Cov[\alpha_i, x_{i,t}]$ may not be 0.

Initial condition $y_{i,0}$ is observed.

Mundlak (1978), Chamberlain (1984): Project α_i on $\mathbf{X}_i$

$\alpha_i = \bar{\mathbf{x}}_i'\boldsymbol{\delta} + u_i$, $u_i \sim N[0,\sigma_u^2]$

Implies a random effects model:

$y_{i,t} = 1[\mathbf{x}_{i,t}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + u_i + \varepsilon_{i,t} > 0]$, t = 1,...,T; i=1,...,N

Problems with Dynamic RE Probit

- Assumes $y_{i,0}$ and the effects are uncorrelated
- Assumes the initial conditions are exogenous – OK if the process and the observation begin at the same time, not if different.
- Doesn't allow time invariant variables in the model.
- The normality assumption in the projection.

Example11.do