xtunbalmd: Dynamic Binary Random Effects Models Estimation with Unbalanced Panels

Pedro Albarran* Raquel Carrasco** Jesus M. Carro**

*Universidad de Alicante
**Universidad Carlos III de Madrid

2017 Spanish Stata Users Group meeting

Introduction

- Focus: to deal with the implementation in Stata of estimators for dynamic binary choice correlated random effects (CRE) models when having unbalanced panel data.
- Data often come from unbalanced panels:
  - unbalancedness generated by sample design, as the Monthly Retail Trade Survey (U.S), the Spanish Family Expenditure Survey.
  - unbalancedness generated by the sample selection process, as the PSID (U.S).

Introduction

- CRE approaches are popular among practitioners to control for permanent unobserved heterogeneity in non-linear models like

$$y_{it} = 1\{\alpha y_{it-1} + X'_{it}\beta + \eta_i + \varepsilon_{it} \geq 0\} \ (t = 1, ..., T; i = 1, ..., N) \qquad (1)$$

- Examples: Hyslop (Ecta. 1999), Contoyannis et. al.(*JAE* 2004), Stewart (*JAE* 2007), Akee et. al.(*Am Econ J Appl Econ.* 2010).

- Why are CRE methods popular?
  1. Simplicity
  2. The alternative fixed effect approach suffers from the incidental parameters problem when the time dimension of the panel is small.

Introduction

- CRE approach disadvantages:
  - It imposes parametric assumptions on the conditional distribution of η_i
  - In dynamic models, the *initial conditions problem*: if the start of the sample does not coincide with the start of the stochastic process, the first observation will not be independent of the time invariant unobserved effect.
    - This problem becomes particularly relevant when having unbalanced panels.
- Solutions proposed to address the *initial conditions problem* (e.g. Heckman, 1981, and Wooldridge, 2005) developed for balanced panels.

Introduction

- Typical "solutions" in empirical work:
  - Ignoring the unbalancedness: only valid under unbalancedness completely at random and no dynamics
  - Extract a balanced panel from the unbalanced sample, so that the existing CRE methods for balanced panels can then be used.
    - For instance, taking the subset of periods constituting a balanced panel for all the individuals: not feasible, efficiency losses.
    - Using only the subset of individuals that stay longer in the panel: not a representative sample, not possible to obtain consistent estimates of the average marginal effects.

Introduction

- We introduce a command "`xtunbalmd`" that performs the estimation of the model for each subpanel separately and obtain estimates of the common parameters across subpanels by minimum distance (MD).
- `xtunbalmd` simplifies the maximum likelihood (ML) estimation in which specific parameters to each sub-panel are jointly estimated with the common parameters of the model, while keeping the good asymptotic properties. It also allows to use the same Stata estimation routines that we would use if we had a balanced panel.
- We also address how to estimate the model using standard built-in commands in Stata by ML (although this can be in some cases computationally cumbersome), and how to estimate models with different assumptions regarding the correlation between the unbalancedness and the individual effects.

The model

- Borrowing the notation from Albarran et. al. (2017), consider the following dynamic binary choice model:

$$y_{it} = 1\left\{\alpha y_{it-1} + X_{it}'\beta + \eta_i + \varepsilon_{it} \geq 0\right\}, \tag{2}$$

$$-\varepsilon_{it} | \, y_i^{t-1}, X_i, \eta_i, S_i \underset{iid}{\sim} N(0,1), \tag{3}$$

  and a random sample of $(Y_i, X_i, S_i) \equiv \{y_{it}, x_{it}, s_{it}\}_{t=1}^{T}$ for N individuals. s_{it} indicates whether individual i is observed in period t.

- Initial conditions problem applies to each first period of observation of the individuals in the sample.

The model

- We write the likelihood function of the sample by specifying the density of the time invariant unobserved heterogeneity, η_i, conditional on the first observation as follows (see Wooldridge, 2005):

$$\Pr\left(S_1'Y_1,\ldots,S_N'Y_N \middle| X_1,\ldots,X_N,S_1,\ldots,S_N\right)$$

$$=\prod_{i=1}^{N}\left[\int_{\eta_i}\prod_{t=t_i+1}^{t_i+T_i-1}\Pr\left(y_{it}|y_{it-1},X_i,S_i,\eta_i\right)h(\eta_i|y_{it_i},X_i,S_i)d\eta_i\right]\Pr\left(y_{it_i}|X_i,S_i\right), \tag{4}$$

where t_i is the first period in which unit i is observed, and T_i is the number of periods we observe for unit i.

$\Pr\left(y_{it}|y_{it-1},M_iX_i,S_i,\eta_i\right)$ is given by

$$\Pr\left(y_{it}=1|y_{it-1},X_i,S_i,\eta_i\right)=\Phi\left(\alpha y_{it-1}+\beta_0+X_{it}'\beta+\eta_i\right). \tag{5}$$

- We specify

$$\eta_i|y_{it_i},X_i,S_i\sim N\left(\pi_{0S_i}+\pi_{1S_i}y_{it_i}+\overline{X}_i'\pi_{2S_i},\sigma^2_{\eta S_i}\right) \tag{6}$$

Implementation

- Previous models can be estimated by Maximum Likelihood (ML).
- For balanced panels, Wooldridge (2005) shows that a simple likelihood can be maximized with standard random-effects probit software ('`xtprobit`' command in Stata).
- However, in our unbalanced case, maximizing the likelihood is cumbersome.
- Simpler implementation: A Minimum Distance estimation.
  - Estimate separately CRE (balanced) probits for each subpanel.
  - Calculate the minimum distance estimates of α and β.

Different assumptions

- **Assumption 1:** Allowing for dependence between S_i and η_i.

This implies that different distributions of the initial conditions and of the unobserved effects for each sub-panel are required.
Following Wooldrdige (2005) we assume

$$\eta_i | y_{it_i}, M_i X_i, S_i \sim N\left(\pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}' \pi_{2S_i}, \sigma^2_{\eta S_i}\right). \tag{7}$$

- **Assumption 2:** Allowing for dependence between t_i and η_i.

The unbalancedness is denoted by two elements: the period each sub-panel starts, t_i, and the number of periods of each sub-panel, T_i (the definition of "subpanel" changes)

Different assumptions

- **Assumption 3:** Independence between S_i and η_i.

Even if we assume that the sample selection process S_i is independent of η_i, the distribution of η_i will be different for each t_i, i.e. it will be:

$$\eta_i | y_{it_i}, M_i X_i, S_i \sim N\left(\pi_{0t_i} + \pi_{1t_i} y_{it_i} + \overline{\overline{M_i X_i}}' \pi_{2t_i}, \sigma^2_{\eta t_i}\right), \tag{8}$$

$\eta_i | y_{it_i}, M_i X_i, S_i$ still has different parameters depending on when each sub-panel starts.

- **Assumption 4:** Allowing for dependence between S_i (or t_i) and η_i only through the mean.

The variance of the distribution of $\eta_i | y_{it_i}, M_i X_i, S_i$ is constant across sub-panels, that is:

$$\eta_i | y_{it_i}, M_i X_i, S_i \sim N\left(\lambda_{0S_i} + \lambda_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}_{2S_i}' \lambda, \sigma^2_{\eta}\right). \tag{9}$$

Estimators

- The contribution to the likelihood function for individual i is given by

$$L_i = \int \prod_{t=t_i+1}^{t_i+T_i} \Phi\left[\left(\alpha y_{it-1} + X_{it}'\beta + \pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{L_i X_i}'\pi_{2S_i} + a\right)(2y_{it} - 1)\right. \tag{10}$$

- The MLE maximizes $\mathcal{L} = \sum_{i=1}^{N} \log L_i$ with respect to the whole set of parameters: $\left(\alpha, \beta', \{\pi_{0j}\}_{j=1}^{J}, \{\pi_{1j}\}_{j=1}^{J}, \{\pi_{2j}\}_{j=1}^{J}, \{\sigma_{\eta j}\}_{j=1}^{J}\right)$
- Maximizing the likelihood is cumbersome and cannot be done using such standard built-in commands.
- Although in theory it is possible to obtain these ML estimates by using the 'gllamm'and/or 'gsem'commands in Stata 13 (or higher), in practice this is not computationally feasible in many cases. See the Albarran et. al. (2017) for details.

Estimators

- We propose an estimation method that allows to use the same routines as when having a balanced panel, while keeping the good asymptotic properties of the MLE.
  - to estimate the model for each subpanel separately, that is, to obtain in a first stage the estimated coefficients for each subpanel by maximizing the likelihood for each subpanel,
  - to obtain estimates of the common parameters across subpanels by MD.
- Practical problem with the MD estimator: potential lack of variability in a specific sub-panel.

Stata implementation

- The command `xtunbalmd` involves two stages:
  - the estimation of the parameters for each sub-panel separately using the Stata command `xtprobit` (that accounts for the initial conditions problem following Wooldridge's approach);
  - the estimation of the common parameters by minimizing the weighted difference between the coefficients obtained in the first stage using a MD approach.
- In addition to the estimated coefficients and their standard errors, `xtunbalmd` also provides estimates and standard error of the marginal effects of the lagged dependent variable.
- The data requirements are basically that the data must contain at least three observations per subpanel.

Options

- The command `xtunbalmd` offers different options, depending on the definition of subpanel, and also depending on the type of correlation between the unbal. structure and the individual effect.

Options

- **Estimation under Assumption 4**

The simplifying assumption that the variance of the conditional distribution of η_i is constant across sub-panels, makes the implementation of the ML estimator easy and feasible. That is, if we assume that

$$\eta_i | y_{it_i}, M_i X_i S_i \sim N\left(\pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{X}_i' \pi_{2S_i}, \sigma_\eta^2\right),$$

ML estimates can be easily obtained by using also the "`xtprobit`" command.

- **Estimation ignoring the unbalancedness and balancing the sample**

The estimation of the models that either ignore or balance the sample can be done very easily using the Stata `xtprobit` command, under the solution proposed by Wooldridge (2005) to solve the initial conditions problem.

Some results

- Monte Carlo experiments and an empirical illustration show that our proposed estimation approaches perform better both in terms of bias and RMSE than the approaches that ignore the unbalancedness or that balance the sample.
- Both the ML and the MD estimators have comparative advantages and disadvantages. Its computational simplicity leads us to favor the MD approach.
  - when estimating the model by ML we make an efficient use of all the observations in the sample, but estimating this model is computationally cumbersome and takes a lot of time because all parameters are jointly estimated: The MLE can take between 150 and 1,600 times more computing time than the MD, depending on the number of periods and subpanels.
  - the MD estimation is much faster. Although we face a potential problem of lack of variability in certain sub-panels,the percentage of simulations that achieved convergence for the MD estimator is very high.

Some results: Empirical Illustration: Export market participation

- Data for Spanish manufacturing firms, the Business Strategies Survey (*Encuesta sobre Estrategias Empresariales*, ESEE).
- Annual data for the period from 1990 to 1999.
- Final sample: unbalanced panel of 1,807 firms and 12,683 observations.
- The comparison between the sets of estimates presented in the empirical application emphasizes the point that different individuals behave differently due to the heterogeneity in the distribution of the unobservables across subpanels. It also reveals the importance of accounting for it to give a proper estimate of the marginal effect of the explanatory variables in a dynamic non-linear model. Example

Table: Unbalancedness structure of the total sample

Subpanel	Number of firms	Pattern by year							
		1990	1991	1992	1993	1994	1995	1996	19
$S = 1$	143	x	x	x	.	.	.	.	.
$S = 2$	100	x	x	x	x	.	.	.	.
$S = 3$	102	x	x	x	x	x	.	.	.
$S = 4$	66	x	x	x	x	x	x	.	.
$S = 5$	63	x	x	x	x	x	x	x	.
$S = 6$	48	x	x	x	x	x	x	x	x
$S = 7$	79	x	x	x	x	x	x	x	x
$S = 8$	699	x	x	x	x	x	x	x	x
$S = 9$	65	.	x	x	x	x	x	x	x
$S = 10$	34	.	.	x	x	x	x	x	x
$S = 11$	37	.	.	.	x	x	x	x	x
$S = 12$	34	.	.	.	.	x	x	x	x
$S = 13$	91	.	.	.	.	.	.	x	x
$S = 14$	246	.	.	.	.	.	.	.	x
$S = 1$ to 14	1,807								
$S = 15$	16	.	.	.	.	.	.	x	x

Table: Estimated Average marginal effects of Lagged Export.

	Bal. Units (1)	Ignore Unbal. (2)	Unbal. MD (3)	Test (2)
Total sample	0.2423	0.2351	0.2776	
	(0.0290)	(0.0234)	(0.0254)	
Subsample, by age††				
$Age < 12$	0.2590	0.2528	0.3181	
	(0.0313)	(0.0251)	(0.0290)	
Age 12-24	0.2735	0.2573	0.2994	
	(0.0314)	(0.0250)	(0.0266)	
$Age > 24$	0.2121	0.2032	0.2307	
	(0.0268)	(0.0212)	(0.0234)	
Subsample, by I.C.				
$\text{Export}_{t_i} = 1$	0.1640	0.1808	0.2064	
	(0.0257)	(0.0209)	(0.0234)	
$\text{Export}_{t_i} = 0$	0.2811	0.2811	0.3391	
	(0.0269)	(0.0269)	(0.0287)	
Subpanels $S \neq 8$		0.2358	0.3267	*
		(0.0236)	(0.0328)	

Table: Estimated Average marginal effects of Lagged Export. By Subpanels

	Bal. Units (1)	Ignore Unbal. (2)	Unbal. MD (3)	Test of Diff. (2) vs (3)
Subpanels				
$S = 1$		0.2414	0.2903	
		(0.0245)	(0.0904)	
$S = 2$		0.2338	0.4380	***
		(0.0239)	(0.0442)	
$S = 3$		0.2470	0.4144	**
		(0.0247)	(0.0776)	
$S = 4$		0.2108	0.2539	
		(0.0218)	(0.1033)	
$S = 5$		0.2340	0.3477	
		(0.0239)	(0.0732)	
$S = 6$		0.2230	0.1095	***
		(0.0222)	(0.0209)	
$S = 7$		0.2182	0.3441	***
		(0.0223)	(0.0477)	
$S = 8$	0.2423	0.2336	0.2413	