2020-02

Working paper. Economics

ISSN 2340-5031

USING STATA TO ESTIMATE DYNAMIC CORRELATED RANDOM EFFECTS PROBIT MODELS WITH UNBALANCED PANELS

Pedro Albarrán, Raquel Carrasco and Jesús M. Carro

uc3m | Universidad Carlos III de Madrid

Serie disponible en http://hdl.handle.net/10016/11

Web: http://economia.uc3m.es/

Correo electrónico: departamento.economia@eco.uc3m.es

Using Stata to Estimate Dynamic Correlated Random Effects Probit Models with Unbalanced Panels*

Pedro Albarran[a], Raquel Carrasco[b], and Jesus M. Carro[† b]

[a]Fundamentos del Análisis Económico (FAE), Universidad de Alicante
[b]Department of Economics, Universidad Carlos III de Madrid

7th April, 2020

Abstract

This paper implements the estimation of dynamic probit correlated random effects (CRE) models with unbalanced panel data. The type of models we consider include a lag of the endogenous variable and other explanatory variables that are strictly exogenous. We introduce a Stata package, **xtprobitunbal**; this command estimates these models allowing for the unbalancedness process to be correlated with the time-invariant unobserved heterogeneity. It reduces the computational burden of the maximum likelihood (ML) estimation, while keeping its good asymptotic properties. We also introduce the command `mgf_unbal` to compute the marginal effects of the variables of the model and its standard errors. Finally, we study the estimation of CRE unbalanced panel data probit models by ML estimation and under more restrictive assumptions than the ones considered by **xtprobitunbal**, discussing the main problems to implement them.

JEL Classification: C23, C25
Keywords: unbalanced panels, correlated random effects, dynamic non-linear models, Stata

*We are grateful to Ricardo Mora and participants at the Spanish Stata Users Group meeting 2017 for helpful comments on this work. All remaining errors are our own. The authors gratefully acknowledge research funding from the Spanish Ministry of Education, Grants ECO2017-87069-P and RTI2018-095231-B-I00.

†Corresponding author. E-mail: jcarro@eco.uc3m.es

1 Introduction

There are important reasons why it is necessary to have specific econometric software for dealing with unbalancedness in the estimation of non-linear dynamic panel data models.[1] As pointed out by Albarran, Carrasco and Carro (2019), using software that just ignores the unbalancedness and treats the data as if they were balanced produces inconsistent estimates of the parameters even if the unbalancedness process is completely at random. Taking a subsample to "make the sample balanced" also presents important caveats. For instance, using the subset of individuals that are observed over the same periods implies an endogenous selection of the sample and, therefore, it is not possible to obtain consistent estimates of the average marginal effects unless the unbalancedness is independent of the individual effects. Also balancing the sample using the subset of periods at which all individuals are observed is in many cases infeasible due to the lack of enough number of common periods across individuals and, when feasible, it discards useful information which may imply important efficiency losses.

In Albarran et al. (2019) we propose methods to deal with the unbalancedness structure of the data in the estimation of dynamic non-linear correlated random effects (CRE) models. The type of models we consider include a lag of the endogenous variable and other explanatory variables that are strictly exogenous. We assume the unbalancedness process to be independent of the time-varying shocks but allow it to be correlated with the time-invariant unobserved heterogeneity. We discuss how to address the estimation by maximizing the likelihood function of the whole sample and propose a minimum distance (MD) approach which is computationally simpler and asymptotically equivalent to the maximum likelihood (ML) estimation.

In this paper we present a Stata command, **xtprobitunbal**, that implements the MD estimator proposed for a general dynamic model in Albarran et al. (2019) to the probit case dealing with the initial conditions problem as in Wooldridge (2005). It reduces the computational burden of the ML estimation while keeping its good asymptotic properties. We also present a post-estimation command, `mgf_unbal`, to compute the marginal effects of the variables of the model, which are the main parameters of interest, and its standard errors. It is important to point out that the procedure we propose can also be implemented using any other software different from Stata with libraries to estimate probit models or to perform maximum likelihood estimation.

As previously explained, existing commands to estimate dynamic CRE probit models, such as `xtprobit` or **redprob** in Stata, should not be used in the presence of unbalanced panel data.[2] This is because when using these commands one either ignores the unbalancedness or extracts a balanced panel from the unbalanced sample which, as previously pointed out, is incorrect. Also, as we will show, the estimation that accounts for the unbalancedness could in principle be performed via joint ML estimation, using commands such as **gsem** or **gllam** in Stata or similar commands for other software. The problem with the ML estimation is that the optimization procedure is cumbersome. The reason is that the likelihood needs to be maximized jointly with respect to a high number of parameters because, due to the unbalancedness, there is a different set of parameters for

[1] Unbalanced panels are often encountered in applied work. The unbalancedness can be driven by sample design, for instance, in the case of rotating panels where the unbalancedness is completely at random, as in the Monthly Retail Trade Survey for the US. In other cases the unbalancedness structure may be related to some of the model's variables, like in panels with attrition as the PSID for the US or the GSOEP for Germany.

[2] `xtprobit` implements the estimation of dynamic probit CRE models under the solution used in Wooldridge (2005) to solve the initial conditions problem, while **redprob** implements the estimation using the Heckman (1987)'s approach; see Stewart (2006). Notice that the initial conditions problem is exacerbated when the panel is unbalanced because it affects to each first period of observation in the data set.

each subpanel.

The rest of the paper is structured as follows. In Section 2, we present the model. In Section 3, we describe the syntax of `xtprobitunbal` and `mgf_unbal` and illustrate their use through an example using simulated data calibrated to the estimates presented in the empirical application in Albarran et al. (2019). Section 4 discusses the implementation of the ML estimation with existing Stata commands and the problems it presents. In Section 5 we present the Stata codes that can be used to estimate models under more restrictive assumptions than the ones considered by the `xtprobitunbal` command, as well as models that account for the initial conditions problem by using Heckman (1987)'s approach. Finally, Section 6 concludes.

2 The model

Borrowing the notation from Albarran et al. (2019), consider the following dynamic binary choice model:

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

$$-\varepsilon_{it} |\, y_i^{t-1}, X_i, \eta_i, S_i \underset{iid}{\sim} N(0,1), \tag{2}$$

and a random sample of $(Y_i, X_i, S_i) \equiv \{y_{it}, x_{it}, s_{it}\}_{t=1}^{T}$ for N individuals. y_{it} is the outcome, X_{it} is a row vector of dimension K of covariates. s_{it} indicates whether y_{it} and X_{it} for individual i are observed. η_i denotes the vector of permanent unobserved heterogeneous characteristics, and ε_{it} are period-specific disturbances that are assumed to be independent and identically distributed across both $i = 1, \ldots, N$ and $t = 1, \ldots, T$ with known distribution.

We assume that ε_{it} is independent of η_i and X_i. This means that we consider models where X are strictly exogenous covariates with respect to the period-specific unobservables, ε, but they can be correlated with the time-invariant unobservables, η_i. We also assume that ε is conditionally independent of the sample selection process S_i that produces the unbalancedness. However, note that this assumption does not restrict the relation between S_i and (η_i, X_i). Therefore, although we do not consider an endogenous selection process with respect to the period-specific disturbances, we allow S_i to be correlated with the unobserved permanent characteristics η_i.

We consider panels for which all the observations for unit i are consecutive. Let M_i be the $(T_i \times T)$ matrix that select the set of X_i that we observe, that is, $M_i X_i = \left(X_{it_i}^{\top}, \ldots, X_{it_i+T_i-1}^{\top}\right)^{\top}$, 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. We denote by J the number of different S_i sequences that we have in the total panel. We refer to the sub-set of units with the same sequence $S^{(j)}$ as "sub-panel" j, $j = 1, \ldots, J$. In other words, subpanel j contains all the individuals i such that $S_i = S^{(j)}$. Finally, we consider panels where N is large and T and J are small relative to N.

The probability of a given random sample of N unit observations is

$$\Pr\left(S_1^{\top}Y_1, \ldots, S_N^{\top}Y_N \middle|\, X_1, \ldots, X_N, S_1, \ldots, S_N\right) = \prod_{i=1}^{N} \Pr\left(S_i^{\top}Y_i | M_i X_i, S_i\right). \tag{3}$$

For each $i = 1, \ldots, N$,

$$\begin{aligned}\Pr(s_{i1}y_{i1}, \ldots, s_{iT}y_{iT}|M_iX_i, S_i, \eta_i) = \\ &= \prod_{t=1}^{T} \Pr\left(y_{it}|s_{it-1}y_{it-1}, M_iX_i, S_i, \eta_i\right)^{s_{it}s_{it-1}} \Pr\left(y_{it}|M_iX_i, S_i, \eta_i\right)^{s_{it}(1-s_{it-1})} \\ &= \prod_{t=t_i+1}^{t_i+T_i-1} \Pr\left(y_{it}|y_{it-1}, M_iX_i, S_i, \eta_i\right) \Pr\left(y_{it_i}|M_iX_i, S_i, \eta_i\right), \qquad (4)\end{aligned}$$

We can write $\Pr\left(S_i^\top Y_i|M_iX_i, S_i\right)$ in equation (3) by making a distributional assumption about η_i conditional on the initial period observation:

$$\left[\int_{\eta_i} \prod_{t=t_i+1}^{t_i+T_i-1} \Pr\left(y_{it}|y_{it-1}, M_iX_i, S_i, \eta_i\right) h(\eta_i|y_{it_i}, M_iX_i, S_i)d\eta_i\right] \Pr\left(y_{it_i}|M_iX_i, S_i\right), \qquad (5)$$

where, from the model equations in (1) and (2), $\Pr\left(y_{it}|y_{it-1}, M_iX_i, S_i, \eta_i\right)$ is

$$\Pr\left(y_{it} = 1|y_{it-1}, M_iX_i, S_i, \eta_i\right) = \Phi\left(\alpha y_{it-1} + \beta_0 + X_{it}^\top\beta + \eta_i\right). \qquad (6)$$

To deal with the initial conditions problem that arises in dynamic models under the CRE framework, as in Wooldridge (2005) we assume

$$\eta_i|y_{it_i}, M_iX_i, S_i \sim N\left(\pi_{0S_i} + \pi_{1S_i}y_{it_i} + \overline{\overline{M_iX_i}}^\top\pi_{2S_i}, \sigma^2_{\eta S_i}\right), \qquad (7)$$

where $\overline{\overline{M_iX_i}}^\top = \frac{1}{T_i-1}\sum_{t=t_i+1}^{t_i+T_i} x_{it}$; see Rabe-Hesketh and Skrondal (2013). Notice that the initial conditions problem becomes particularly relevant with unbalanced panels because it applies to each first period of observation of the individuals in the sample.

Alternatively, we can consider a model in which the unobserved effect is integrated out by specifying the density for the first observation in each sub-panel conditional on the unobserved effect, $\Pr\left(y_{it_i}|M_iX_i, S_i, \eta_i\right)$, and the density of the unobserved effect. Then, we can write the probability $\Pr\left(S_i^\top Y_i|M_iX_i, S_i\right)$ in (3) as

$$\begin{aligned}\Pr\left(s_{i1}y_{i1}, \ldots, s_{iT}y_{iT}|M_iX_i, S_i\right) = \int_{\eta_i} \prod_{t=t_i+1}^{t_i+T_i} \Pr\left(y_{it}|y_{it-1}, M_iX_i, S_i, \eta_i\right) \times \\ \Pr\left(y_{it_i}|M_iX_i, S_i, \eta_i\right) h(\eta_i|M_iX_i, S_i)d\eta_i \qquad (8)\end{aligned}$$

To solve the initial conditions problem in this case we can follow Heckman (1987)'s approach and use for the first observation the same parametric form as the conditional density for the rest of the observations:

$$\begin{aligned}\Pr\left(y_{it_i} = 1|X_i, S_i, \eta_i\right) &= \Pr\left(y_{it} = 1|X_{it}, S_i, \eta_i, s_{it-1} = 0, s_{it} = 1\right) \\ &= \Phi\left(\delta_{0S_i} + X_{it_i}^\top\delta_{S_i} + \mu_{S_i}\eta_i\right), \qquad (9)\end{aligned}$$

where we have different distributions for each value of S_i because we allow for correlation between S_i and η_i.

For the density of the unobserved effect, $h(\eta_i|X_i, S_i)$, we can follow Chamberlain (1980) to allow for correlation between the individual effect and the explanatory variables:

$$\eta_i|X_i, S_i \sim N\left(\overline{X}_i^\top\beta_{\eta S_i}, \sigma^2_{\eta S_i}\right), \qquad (10)$$

where $\overline{X}_i$ contains the within-means of the time-varying explanatory variables. Notice that (10) allows for correlation between the sample selection process, S_i, and the permanent unobserved heterogeneity η_i.

Albarran et al. (2019) show that both approaches to write the likelihood function, (5) and (8), have similar performance. As a consequence of that, the command we develop is based only on (5) because for the distributions and functional forms considered, it is faster to compute. Nonetheless, Section 4 includes a discussion on how to estimate the model both under the Heckman's approach and Wooldridge's proposal using the ML estimation procedure and the main problems in doing so. Section 5.2 explains why when the unbalancedness is independent of η_i the specification using Heckman's approach is easier to estimate by ML.

2.1 Minimum distance estimator

Model (5) can be estimated by ML. The contribution to the likelihood function for individual i is given by

$$L_i = \int \prod_{t=t_i+1}^{t_i+T_i-1} \Phi\left[\left(\alpha y_{it-1} + X_{it}^{\top}\beta + \pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}^{\top} \pi_{2S_i} + a\right)(2y_{it}-1)\right] \times \frac{1}{\sigma_{\eta S_i}} \phi\left(\frac{a}{\sigma_{\eta S_i}}\right) da. \quad (11)$$

The ML estimator maximizes $\mathcal{L} = \sum_{i=1}^{N} \log L_i$ with respect to

$$\theta \equiv \left(\alpha, \beta^{\top}, \{\pi_{0j}\}_{j=1}^{J}, \{\pi_{1j}\}_{j=1}^{J}, \{\pi_{2j}\}_{j=1}^{J}, \{\sigma_{\eta j}\}_{j=1}^{J}\right)^{\top}.$$

The properties of the ML estimator are well-known, as well as the numerical procedures to obtain it. The problem is that the optimization procedure is cumbersome because the likelihood function should be maximized with respect to a high number of parameters: the vector of common parameters and the set of sub-panel specific parameters. This will typically preclude using standard estimation software and will increase computation time.[3]

We propose a procedure to estimate the model using a MD approach. This procedure allows us to take advantage of the existing routines or estimation programs for balanced panels, while keeping the good asymptotic properties of the ML estimator and reducing its computational burden. The proposal has two steps:

1. Estimate by ML the model for each sub-panel separately using the same standard software as when having balanced panels.[4] That is, we obtain in a first stage $\widehat{\delta} = \left(\widehat{\delta}_1^{\top}, \widehat{\delta}_2^{\top}, \ldots, \widehat{\delta}_J^{\top}\right)^{\top}$ by maximizing $\mathcal{L}_j = \sum_{i \in \{i:S_i = S^{(j)}\}} \log L_i$ for each subpanel $j = 1, \ldots, J$.

2. Obtain the estimates of the common parameters across subpanels by MD.[5] Notice that each $\widehat{\delta}_j^{\top}$ includes two types of parameters: $\widehat{\delta}_j^{[c]}$, the estimates of the parameters that are common across subpanels, and $\widehat{\delta}_j^{[nc]}$, the estimates of the non-common parameters for sub-panel j.

[3] Although in theory it is possible to obtain these ML estimates by using the **gllamm** and/or **gsem** commands in Stata (version 13 or higher), in practice this is not computationally feasible in many cases. See Section 4 for details.

[4] In this stage one can use the existing commands to perform the estimation as in balanced panels, both following the Wooldridge's and the Heckman's approach.

[5] It is important to note that, although computationally feasible, a potential practical problem with the MD estimator could be the lack of variability in a specific subpanel.

To recover the estimate of the common parameters, we assume that all the $\widehat{\delta}_j^{[c]}$ are estimates of the same common parameters. Therefore the restrictions are

$$h(\theta) = \begin{pmatrix} h_1(\theta) \\ \vdots \\ h_J(\theta) \end{pmatrix} = P\theta.$$

The structural parameters θ can be consistently and efficiently estimated by minimizing the following quadratic form:

$$\widehat{\theta}^{MD} = \arg\min_{\theta} Q(\theta) = \left[\widehat{\delta} - h(\theta)\right]^{\top} V^{-1} \left[\widehat{\delta} - h(\theta)\right]. \tag{12}$$

The solution to the minimization of this quadratic form is

$$\widehat{\theta}^{MD} = \left[P^{\top} V^{-1} P\right]^{-1} P^{\top} V^{-1} \widehat{\delta}, \tag{13}$$

where V is replaced by a consistent estimator obtained in the first step; see Albarran et al. (2019), for details.

2.2 Average marginal effects

The average marginal effects (AMEs), which are ultimately the parameters of interest, are based on

$$E[\Phi\left(\alpha y_{it-1} + X_{it}^{\top}\beta + \eta_i\right)], \tag{14}$$

Using that $\eta_i = \pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}^{\top} \pi_{2S_i} + \xi_i$ and following Wooldridge (2005), expression (14) becomes

$$E\left[\Phi\left(\frac{\alpha y_{it-1} + X_{it}^{\top}\beta + \pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}^{\top} \pi_{2S_i}}{\sqrt{1 + \sigma^2_{\eta S_i}}}\right)\right], \tag{15}$$

where this expectation is taken with respect to the distribution of the covariates conditional on the unbalancedness structure, $\{S^{(1)}, \ldots, S^{(J)}\}$.

The AME for a continuous regressor is the derivative of (15) with respect to that regressor, and the AME for a discrete regressor is the difference in expression (15) for a unitary change in the regressor. It is worth noting that previous expression depends on the correlation between the unbalancedness and this individual effect. Therefore, when this correlation is neglected, the estimates of the AMEs will be biased.

The estimated AME, $\widehat{AME}$, can be simply obtained by replacing the population expectation in (15) with the sample mean. For instance, the $\widehat{AME}$ for the lagged dependent variable is:

$$\begin{aligned} \widehat{AME}_{y_{t-1}} = & \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\frac{\widehat{\alpha} + X_{it}^{\top}\widehat{\beta} + \widehat{\pi}_{0S_i} + \widehat{\pi}_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}^{\top} \widehat{\pi}_{2S_i}}{\sqrt{1 + \widehat{\sigma}^2_{\eta S_i}}}\right) \\ & - \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\frac{X_{it}^{\top}\widehat{\beta} + \widehat{\pi}_{0S_i} + \widehat{\pi}_{1S_i} y_{it_i} + \overline{\overline{M_i X_i}}^{\top} \widehat{\pi}_{2S_i}}{\sqrt{1 + \widehat{\sigma}^2_{\eta S_i}}}\right). \end{aligned} \tag{16}$$

Finally, the standard errors of the $\widehat{AMEs}$ are computed using the Delta method.

3 Syntax and implementation of the MD procedure

In this section we describe how users can implement the MD estimation procedure previously described using the package **xtprobitunbal** in Stata. This includes the estimation command `xtprobitunbal` and the post-estimation command `mgf_unbal` for marginal effects computation. It can be installed from the Statistical Software Components (SSC)[6] archive by typing on the Stata command line the following code:

```
. ssc install xtprobitunbal, all replace
```

Two files are also downloaded with the code above: the dataset `exportunbal.dta` and `exportunbal.do`, a Stata code script with examples. These are intended to familiarize users with the commands.

The basic syntax to obtain the parameter estimates is as follows:

```
xtprobitunbal depvar indepvars [if] [in], meansvar(varlist) ///
    [gensubp(varname) indep niterat(#) quatp(#)]
```

The lag value of `depvar` is included by default. The list `indepvars` is a non-optional variable list which includes the exogenous covariates of the model. Moreover, `meansvar(varlist)` should specify a list of variables the means of which will be included in the reduced form for the initial condition equation.

Being a Stata `[xt]` command, it requires the command `xtset` to be run in advance to declare the panel structure, using a panel variable (and, optionally, a time variable). By default, a subpanel is defined as each of the different time patterns in the data set, defined by the first and last time period in which an individual is observed.[7] In this case, observations for which the first and final periods are the same belong to the same subpanel. Under this setting, the `xtprobitunbal` command allows for the unbalancedness process to be correlated with the time-invariant unobserved heterogeneity.

Alternatively, users can set the option `indep`. In such a case, the command `xtprobitunbal` consider a subpanel to be defined only by the initial period; thus, individuals with the same initial period belongs to the same subpanel. If subpanels are defined in this manner, `xtprobitunbal` estimates the econometric model under the underlying assumption that the unbalancedness is independent of the initial condition; see Section 5.2.

Note that the econometric method requires each subpanel to contain at least three time observations per individual and to have enough variation for the estimation of the correlated random effects model for each subpanel. These conditions also applies to the ML estimator. If any of these requirements are not met in a given subpanel, this will be excluded in the first stage of the procedure. When this happens, the command output informs the user and the procedure continues with all the remaining valid subpanels.

There are three additional minor options. On the one hand, `gensubp` allows to specify a variable name where the subpanel index is stored. On the other hand, `niterat` and `quatp` are options controlling the correlated random effects estimation (number of iterations and number of quadrature points, respectively).

After `xtprobitunbal`, users will typically run the command `mgf_unbal` to obtain the marginal effect of an explanatory variable in the model, either the lag or a control variable; note that only one variable is admitted at a time. The syntax for this post-estimation command is as follows:

```
mgf_unbal [if] [in], dydx(string) [val0(#) val1(#)]
```

[6] See https://ideas.repec.org/s/boc/bocode.html

[7] These patterns are shown in Stata by the command `xtdescribe`.

In the mandatory option `dydx`, users specify both the variable whose marginal effect will be calculated and whether this is a discrete or a continuous variable. Three possible types of input are accepted in `dydx` for `string`:

- `lag`: to compute the effect of a discrete change (from 0 to 1) of the lagged dependent variable.
- `d.varname`: to compute the effect of a discrete change of the variable `varname`.
- `c.varname`: to compute the marginal effect of an infinitesimal change of the continuous variable `varname`.

In the case of discrete changes (`d.varname`), the options `val0(#)` and `val1(#)` can also be specified. The command will compute the marginal effect of a discrete change in the variable `varname` when it changes from the value set in `val0` to the value set in `val1`. Defaults values are `val0(0)` and `val1(1)`.

3.1 Example

In this section, we illustrate the implementation of `xtprobitunbal` and `mgf_unbal` by estimating a model for firms' export market participation decision. We use the data set `exportunbal.dta` that is available when installing the package, jointly with the companion Stata example script `exportunbal.do`.[8] These are simulated data, calibrated to the estimates presented in Albarran et al. (2019) using data for Spanish manufacturing firms from the Business Strategies Survey (*Encuesta sobre Estrategias Empresariales*, ESEE) for the period 1990 to 1999. The sample consists of an unbalanced panel with 14 different subpanels of 1,807 firms and 12,683 observations. The dependent variable ($Export_{it}$) is a dummy equal to 1 if the $i-th$ firm exported in year t. In addition to $Export_{it-1}$, the explanatory variables of the model are: firm's size (number of employees/100), share of medium skilled workers (workers with a high school degree), firm's age (years since firm creation/10), and a time trend.

The `xtprobitunbal` command returns

```
. xtprobitunbal export size trend med_skill age, meansvars(size med_skill)

Minimun Distance Estimation of common parameters
for Correlated Random Effects dynamic probit

Number of observations =    10876              Number of groups  =       1807

Number of sub-panels   =       14              Log likelihood    =   -2476.06
------------------------------------------------------------------------------
      export |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Common       |
      export |
         L1. |   1.528326   .0665742    22.96   0.000     1.397843    1.658809
             |
        size |   .1778051    .087053     2.04   0.041     .0071844    .3484258
       trend |   .1082449   .0119901     9.03   0.000     .0847447     .131745
   med_skill |   .0012457   .0044586     0.28   0.780    -.0074929    .0099844
         age |   .0200351   .0174813     1.15   0.252    -.0142276    .0542977
```

[8] The following and additional examples of code can be found in this script.

```
------------------------------------------------------------------------------

Subpanels actually used in estimation:
   1 2 3 4 5 6 7 8 9 10 11 12 13 14
(Variable _subpanel_xtprobitunbal contains the subpanel index)
```

The heading of the output display provides basic information on the model, the estimation procedure, and the sample and number of subpanels used in the estimation. It presents the coefficient estimates of the common parameters.

By default, `xtprobitunbal` stores the estimates of the model parameters, both the common and the specific parameters for each subpanel, as well as the mean and the variance of the sub-panel specific fixed effects in a variable named `e(finalB)`. These values can be used for subsequent analysis, for example to analyse to what extent the distributions of the fixed effects differ by subpanel. Thus, using the values stored in `e(finalB)`, we can draw a graph with the distribution for the fixed effects by subpanel as specified in equation (7), abstracting from $\overline{\overline{M_i X_i}}^{\top}$. We obtain

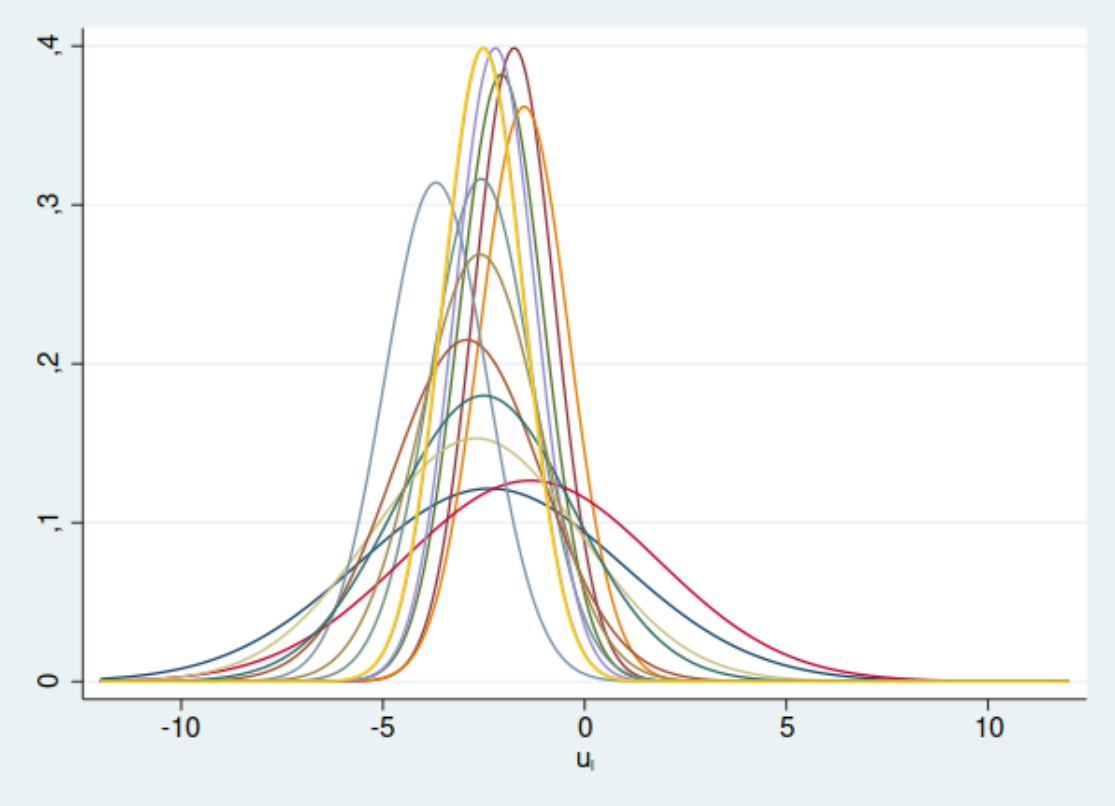

Figure 1: Density functions of u_i for each subpanel.

We can see in the graph that there are substantial differences in the variances of the individual effects by subpanel, although the means are more similar. We can also be interested in analysing to what extent the aggregate distribution can be approximated by a normal distribution. Again, using the information stored in `e(finalB)`, one can present the values for the Skewness and Kurtosis statistics for the simulated variable η_i. We obtain a Skewness equal to 1.66 and a Kurtosis equal to 30.26. These values are far from the values for a Normal distribution, 0 and 3 respectively. A normality test confirms this conclusion since the p value for the test is smaller than 0.0001.

The main parameters of interest in this type of non-linear models, the marginal effects of the variables, can be obtained by using the post-estimation command `mgf_unbal`. In

particular, for the marginal effect of the lagged dependent variable we obtain

```
. mgf_unbal, dydx(lag)

Computing marginal effect of a discrete change from 0 to 1 for the lag of the
 dependent variable export.
       panel variable:  id (unbalanced)
        time variable:  time, 1990 to 1999
                delta:  1 unit

Number of observations =    10876               Number of groups  =       1807

------------------------------------------------------------------------------
             |             Delta-method
             |     AME       Std. Err.      z      P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
    L.export |  .2870526     .0213657     13.44    0.0000     .2451765   .3289286
------------------------------------------------------------------------------
```

The heading of the output display provides information on the marginal effect that has been calculated, and basic information on the sample. As another example, the marginal effect of a discrete change of a variable different than the lagged dependent variable can also be computed. For instance, for the effect of a discrete change in age from the value 2 to the value 3 we obtain:

```
. mgf_unbal, dydx(d.age) val0(2) val1(3)

Computing marginal effect of a discrete change from 2 to 3 for age.
    panel variable:  id (unbalanced)
     time variable:  time, 1990 to 1999
             delta:  1 unit

Number of observations =    10876               Number of groups  =       1807

------------------------------------------------------------------------------
          |             Delta-method
          |     AME       Std. Err.      z      P>|z|    [95% Conf. Interval]
----------+----------------------------------------------------------------
      age |   .002598     .0022833     1.138    0.2552   -.0018771    .0070732
------------------------------------------------------------------------------
```

4 Discussion on ML estimation

The model considered in this paper, that allows for correlation between the unbalancedness structure and the individual effect, can be also estimated by ML. To obtain the ML estimates one has to write the expression for the likelihood function in any likelihood maximizing program. Or in some software, as in Stata, it is also possible to use the commands **gsem** or **gllamm** to obtain the ML estimates. Nonetheless, any of these alternatives are computationally very cumbersome and more difficult to implement.

In this section, we illustrate this point. In particular, we present how the **gsem** and **gllamm** commands in Stata can be used to estimate these type of models. Suppose that the likelihood function to be maximized is the one considered by the `xtprobitunbal` command in equation (11). Suppose also to simplify that we have two sub-panels defined

by the indicator j, $j = 1, 2$. Before calling **gllamm** we need to specify that there are two conditional distributions for the random effects, one for each sub-panel. Therefore, we need to generate first two different constants, `const_1` and `const_2`, and two initial conditions, `y0_1` and `y0_2`, to be included in the main equation. Then, the **gllamm** command is specified as follows:

```
eq etai_1: const_1
eq etai_2: const_2
gllamm y l.y x const_1 const_2 y0_1 y0_2 mx_1 mx_2, ///
    i(id) nrf(2) eqs(etai_1 etai_2)                 ///
    nip(#) fam(binom) link(probit)                  ///
    adapt trace iterate(#) nocorrel noconst
```

where `x` is a covariate and `mx_1` and `mx_2` are the means of the covariate interacted with `const_1` and `const_2`, respectively; of course, additional covariates could be included in a similar manner. The `nrf(2)` option indicates that there are two random effects and the equations `eq etai_1` and `eq etai_2` specify the variables associated to them. The `nocorrel` option specifies zero correlation between the two random effects.

Notice that as the number of sub-panels increases the implementation of this command becomes infeasible. Similar problems can be found when using the **gsem** command. In this case, we have to specify three equations: one for the main dynamic model and two for the initial conditions. We have also to generate the variables $x0\_1$ and $x0\_2$ defined as the values of the regressor for the first time period for subpanel 1 and 2, respectively. Then the **gsem** command is executed by coding

```
xi: gsem( y     <- l.y x I[id] J[id], probit) ///
        ( I[id] <- x0_1, probit)              ///
        ( J[id] <- x0_2, probit)
```

Notice again that, as the number of sub-panels increases, the number of equations to be included in the command also increases. Therefore, the estimation procedure becomes increasingly complex and it often fails to achieve convergence.

Suppose that instead of maximizing the likelihood function in (11) we follow the Heckman's approach and approximate the joint probability of the full observed y sequence as follows:

$$L_i = \int_{\eta_i} \Phi\left(\delta_{0S_i} + X_{it_i}^{\top}\delta_{S_i} + \mu_{S_i}\eta_i\right)(2y_{it_i} - 1)\times \left\{\prod_{t=t_i+1}^{t_i+T_i} \Phi\left[\left(\alpha y_{it-1} + \beta_0 + X_{it}^{\top}\beta + \eta_i\right)(2y_{it} - 1)\right]\right\} h(\eta_i|X_i, S_i)d\eta_i \quad (17)$$

In this case, the generalization of the **gsem** command basically consists on specifying one initial condition equation different for each sub-panel, while the dynamic equation for the rest of observations is common to all the individuals. For instance, for the two sub-panels case the command is specified as follows:

```
xi: gsem( y    <- l.y x I[id], probit)  ///
        ( y0_1 <- J[id] x0_1, probit)   ///
        ( y0_2 <- K[id] x0_2, probit)
```

Notice that a different latent variable should be included in each equation to ensure that the unobserved effect follows a different distribution in each sub-panel. Again, as the number of sub-panels increases the number of equations to include in the command also increases.

The implementation of the **gllamm** command for this model is also difficult. Following with the previous example of two sub-panels, we need to generate not only `const_1` and `const_2`, but also two different dummy variables for the initial conditions, `d0_1` and `d0_2`, taking in this case the value one if the observation corresponds to the first (or second) subpanel and the first period, and 0 otherwise. Then, the **gllamm** command is specified by coding

```
eq etai_1: const_1 d0_1
eq etai_2: const_2 d0_2
gllamm y l.y x1 d0_1 d0_2 x0_1 x0_2 mx_1 mx_2, ///
    i(id) nrf(2) eqs(etai_1 etai_2) ///
    nip(#) fam(binom) link(probit)  ///
    adapt trace iterate(#) nocorrel
```

We have performed some simulations (available upon request) for a model without covariates which show that, as expected, the behaviour of the MD and ML estimators is very similar, in terms of both the estimated parameters and the marginal effects. Nonetheless, the computation time of the ML estimation can be between 150 and 1,600 times greater than that of the MD, depending on the number of periods and subpanels. And this time will further increase when adding covariates.

5 Stata implementation under other assumptions

In this Section we explain how to estimate CRE panel data probit models under more restrictive assumptions than the considered by `xtprobitunbal` by using standard RE probit Stata software. We consider two simplifying assumptions: (i) imposing that the variance of the conditional distribution of η_i is constant across sub-panels, and (ii) assuming that the unbalancedness is independent of the individual effect. Under any of these assumptions both the estimates obtained with **xtprobitunbal** and the estimates presented in next subsections are consistent, but the latter are more efficient. However, if those assumptions are not correct, **xtprobitunbal** still produces consistent estimates whereas the others don't.

5.1 Constant conditional variance of η_i

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

$$\eta_i | y_{it_i}, M_i X_i S_i \sim N\left(\pi_{0S_i} + \pi_{1S_i} y_{it_i} + \overline{X}_i^{\top} \pi_{2S_i}, \sigma_\eta^2\right), \tag{18}$$

ML estimates can be easily obtained by using the `xtprobit` command. One just have to generate as many constants as different subpanels we have and different initial conditions for each subpanel. For instance, for a two sub-panels case without additional regressors, the Stata code would be the following:

```
xtprobit y l.y const_1 const_2 y0_1 y0_2, re iter(#) intpoints(#)
```

where `l.y` is the lagged of the dependent variable, `const_1` and `const_2` are the two intercepts, and `y0_1` and `y0_2` are the initial conditions for each subpanel.

Notice, however, that this assumption is particularly unrealistic when there is correlation between η_i and X_i. The reason is that even under the assumption that S_i is independent of η_i, if the individuals are not observed the same number of periods one would expect the variance of the distribution of $\eta_i | M_i X_i$ to be a function of the number of periods observed and, therefore, to be different for each S_i. Actually, in our example in Section 3.1 the variances change substantially across subpanels.

5.2 Unbalancedness independent of the individual effect

One could assume that S_i is independent of η_i, so that $h(\eta_i|M_iX_i, S_i) = h(\eta_i|M_iX_i)$. This assumption would be relevant, for instance, when having rotating panels. Let's also assume that $h(\eta_i|M_iX_i)$ is a function common to all S_i, so that its value changes only as the values of the $X^\top s$ at which it is evaluated change (but not as a function of the specific periods at which X_i is observed). Notice that even under these assumptions we do not obtain a conditional distribution of η_i common to all sub-panels. That is, $h(\eta_i|y_{it_i}, M_iX_i, S_i)$ will be different for each t_i. In particular, it will be

$$\eta_i|y_{it_i}, M_iX_i, S_i \sim N\left(\pi_{0t_i} + \pi_{1t_i}y_{it_i} + \overline{\overline{M_iX_i}}^\top \pi_{2t_i}, \sigma^2_{\eta t_i}\right), \tag{19}$$

unless the process is not dynamic or it is in its steady state since $t = 0$, or y_{t_i} comes from the same exogenous distribution for all units and t_i.

As can be seen in (19) $\eta_i|y_{it_i}, M_iX_i, S_i$ still has different parameters depending on the period each subpanel starts even under independence of the unbalancedness from η_i. This implies again a complicated structure of the likelihood function and, therefore, the computation of the ML estimator in this case is not much simpler than in the general situation without independence. Therefore, we can use `xtprobitunbal` to avoid the joint ML estimation. For that, we only need to define properly the subpanels using the `indep` option to indicate that they only differ in the initial period observation but not in their duration.

It is interesting to point out that if we write the likelihood function by specifying the density of the first observation conditional on the unobserved effect to deal with the initial conditions problem, as in (17), the ML estimation under independence is simplified and using the **gsem** or the **gllamm** commands becomes feasible. Notice that the difference with respect to the correlated case is that there is only one common distribution for the unobserved effects in all subpanels. In this case, the likelihood function to be maximized is

$$L_i = \int_{\eta_i} \Phi\left(\delta_{0S_i} + X^\top_{it_i}\delta_{S_i} + \mu_{S_i}\eta_i\right)(2y_{it_i} - 1)\times \left\{\prod_{t=t_i+1}^{t_i+T_i} \Phi\left[\left(\alpha y_{it-1} + \beta_0 + X^\top_{it}\beta + \eta_i\right)(2y_{it} - 1)\right]\right\} h(\eta_i|X_i)d\eta_i \tag{20}$$

Then, the **gsem** command is specified as follows:

```
gsem( y    <- l.y x I[id], probit)  ///
    ( y0_1 <- I[id] x0_1, probit)   ///
    ( y0_2 <- I[id] x0_2, probit)
```

The difference with respect to the correlated case is that the same latent variable, $I[id]$, is included in all the equations.

The model based on the likelihood (20) can be also implemented using the **gllamm** command. According to Arulampalam and Stewart (2009), it only requires to specify one equation for the random effect, with one constant, *const*, and two different dummy variables for the initial conditions, $d0\_1$ and $d0\_2$:

```
eq etai: const d0_1 d0_2
gllamm y l.y x1 d0_1 d0_2 x0_1 x0_2 mx_1 mx_2, ///
    i(id) nrf(1) eqs(etai)          ///
    nip(#) fam(binom) link(probit)  ///
    adapt trace iterate(#)
```

6 Conclusion

In this article we have described how to implement a minimum distance estimator for dynamic probit CRE models with unbalanced panels using the **xtprobitunbal** package for Stata. The method allows researchers to obtain consistent estimates of the coefficients of the model and also of the marginal effects, the ultimate parameters of interest, by using an estimation procedure that is computationally more tractable than the ML estimator.

We have illustrated the use of `xtprobitunbal` using an example from Albarran et al. (2019). We have shown the importance of accounting properly for the unbalanced structure of the data by allowing for the specific subpanel fixed-effects to have different distributions. We have found that those distributions do differ across subpanels and that the assumption of aggregate normality is far from being accepted.

We study alternative ways of doing the estimation. As a result of that, we conclude that the only situation in which it is feasible to use existing Stata software for correlated random effects probit models is to impose more restrictive assumptions than the ones considered by **xtprobitunbal**. If one is not willing to impose those assumptions, **xtprobitunbal** represents an useful alternative to the, computationally cumbersome, joint ML estimation.

References

Albarran, Pedro, Raquel Carrasco, and Jesus M Carro, "Estimation of Dynamic Nonlinear Random Effects Models With Unbalanced Panels," *Oxford Bulletin of Economics and Statistics*, 2019, *81* (6), 1424–1441.

Arulampalam, Wiji and Mark B Stewart, "Simplified Implementation of the Heckman Estimator of the Dynamic Probit Model and a Comparison With Alternative Estimators," *Oxford bulletin of economics and statistics*, 2009, *71* (5), 659–681.

Chamberlain, Gary, "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 1980, *47* (1), 225–238.

Heckman, James J, "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in Charles F Manski and Daniel McFadden, eds., *Structural Analysis of Discrete Data With Econometric Applications*, MIT Press Cambridge, MA 1987, pp. 114–178.

Rabe-Hesketh, Sophia and Anders Skrondal, "Avoiding Biased Versions of Wooldridge's Simple Solution to the Initial Conditions Problem," *Economics Letters*, 2013, *120* (2), 346–349.

Stewart, Mark B, "redprob-A Stata Program for the Heckman Estimator of the Random Effects Dynamic Probit Model," 2006. mimeo.

Wooldridge, Jeffrey M, "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models With Unobserved Heterogeneity," *Journal of applied econometrics*, 2005, *20* (1), 39–54.