OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 81, 6 (2019) 0305–9049
doi: 10.1111/obes.12308

Estimation of Dynamic Nonlinear Random Effects Models with Unbalanced Panels*

PEDRO ALBARRAN† RAQUEL CARRASCO‡ and JESUS M. CARRO‡

†Fundamentos del Análisis Económico (FAE), Universidad de Alicante, Alicante, Spain (e-mail: albarran@ua.es)
‡Department of Economics, Universidad Carlos III de Madrid, Getafe, Spain (e-mail: rcarras@eco.uc3m.es, jcarro@eco.uc3m.es)

Abstract

This paper presents estimation methods for dynamic nonlinear models with correlated random effects (CRE) when having unbalanced panels. Unbalancedness is often encountered in applied work and ignoring it in dynamic nonlinear models produces inconsistent estimates even if the unbalancedness process is completely at random. We show that selecting a balanced panel from the sample can produce efficiency losses or even inconsistent estimates of the average marginal effects. We allow the process that determines the unbalancedness structure of the data to be correlated with the permanent unobserved heterogeneity. We discuss how to address the estimation by maximizing the likelihood function for the whole sample and also propose a Minimum Distance approach, which is computationally simpler and asymptotically equivalent to the Maximum Likelihood estimation. Our Monte Carlo experiments and empirical illustration show that the issue is relevant. Our proposed solutions perform better both in terms of bias and RMSE than the approaches that ignore the unbalancedness or that balance the sample.

I. Introduction

The purpose of this paper is to present and evaluate estimation methods for dynamic nonlinear models with correlated random effects (CRE) when the panel data are unbalanced.[1] Unbalanced panels are often encountered in applied work. For example, in large households

JEL Classification numbers: C23, C25.

*We are grateful to César Alonso-Borrego, Wiji Arulampalam, Ricardo Mora, Jeff Wooldridge and participants at the 21st International Panel Data Conference, the IAAE 2015 Annual Conference, the XXXIX SAEe and Econometrics Seminar at Oxford University 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 ECO2012-31358, ECO2015-65204-P, ECO2017-87069-P, MDM 2014-0431 and Comunidad de Madrid, MadEco-CM (S2015/HUM-3444).

[1] The CRE approach has been found useful to estimate nonlinear dynamic models in many cases, because it is not subject to the incidental parameters problem that the fixed-effects (FE) approach suffers and it does not require a large number of periods. Examples of applications using CRE are Hyslop (1999), Contoyannis, Jones and Rice (2004), Stewart (2007) and Akee *et al.* (2010).

panel data sets like the PSID for the U.S. or the GSOEP for Germany, some individuals drop out (potentially non-randomly) of the sample. At a firm level, Compustat and Datastream International also have an unbalanced structure. In other cases, like in the so-called ‘rotating panels’, the unbalancedness is generated by the sample design (for instance, in the Monthly Retail Trade Survey for the U.S., or in the Household Budget Continuous Survey for Spain).

It is well-known how to estimate CRE dynamic nonlinear models with balanced panels. However, the existing estimation methods cannot be in general directly implemented with unbalanced panels. Ignoring the unbalancedness produces inconsistent estimates, as we will discuss. Obtaining a balanced subsample from the unbalanced panel, so that the existing CRE methods for balanced panels could then be used, is also problematic. If we balance the sample by taking a subset of individuals that are observed over the same periods, we are making an endogenous selection of the sample unless the unbalancedness is independent of the individual effects. Another possibility to balance the sample is to take the subset of periods at which all individuals are observed (see Wooldridge, 2005). But this is in some cases infeasible because of the lack of a sufficient number of common periods across individuals and, when feasible, it implies important efficiency losses.

In a dynamic setting under the CRE approach, the so-called ‘initial conditions problem’ arises. Heckman (1981) and Wooldridge (2005) propose solutions to deal with it, but these are developed only for balanced panels. Furthermore, the initial conditions problem is exacerbated when the panel is unbalanced because it affects each of the first period of observation in the data set. This implies that, as we will show, even assuming that unbalancedness is completely at random is not enough to allow us to ignore it in the estimation.[2]

We propose methods to deal with the unbalancedness structure of the data in the estimation of models with lags of the endogenous variable and other explanatory variables that are strictly exogenous. We consider unbalancedness processes that are independent of the time-varying shocks, but allow them to be correlated with the time-invariant unobserved heterogeneity. Therefore, we are not restricted to the case of unbalancedness completely at random. We first discuss how to address the unbalancedness problem by maximizing the likelihood function for the whole sample. This can be computationally cumbersome because specific parameters to each subpanel need to be estimated jointly with the common parameters of the model. We then propose to estimate the model for each subpanel separately and then to obtain estimates of the common parameters across subpanels by minimum distance (MD). This method allows us to use the same estimation routines that we would use if we had a balanced panel, while keeping the good asymptotic properties of the maximum likelihood (ML) estimator for the whole sample.

A simulation study shows that these methods perform well compared to other alternatives both in terms of bias and RMSE. As an empirical illustration, we estimate an export participation equation with dynamic effects using unbalanced data for Spanish manufacturing firms. Our results show that the unbalancedness issue is relevant in practice, and there is evidence of unbalancedness correlated with the unobserved heterogeneity.

[2] This problem also affects RE models assuming that the time invariant unobserved heterogeneity is independent of the time-varying covariates. The CRE setting contains RE models as a particular case.

To the best of our knowledge only Wooldridge (2019) addresses the issue of estimating CRE models with unbalanced panels, but considering only static models. He proposes several strategies for allowing the time invariant unobserved heterogeneity to be correlated with the observed covariates and the selection mechanism for unbalanced panels. However, the assumption of lack of dynamic effects is very restrictive, and the solutions in Wooldridge (2019) cannot be directly extended to dynamic models because the unbalancedness also affects how to deal with the initial conditions problem.

Although unbalanced panels could be seen as a particular case of missing data, the problem we address cannot be solved using the existing literature on panel models with missing data. One strand of this literature relies on missingness at random and on using moment conditions that are valid both with complete and with missing data (e.g. Pacini and Windmeijer, 2015). In our case, the sets of moment conditions (the first-order conditions of the likelihood) under complete data are not valid. The reason is that the likelihood for complete balanced panels does not account for the different initial conditions, or for the potential relation between the unbalancedness and permanent unobserved heterogeneity, as we will show. The other strand of the literature relies on having some variables upon which you can condition to make the missing process conditionally independent of the main model (e.g. Wooldridge, 2007), or on having additional information and assumptions about the missing process (e.g. Bhattacharya, 2008). In contrast with that, we do not assume anything about the relation of the missing process and observable variables, nor have any additional information related to the missing process. Also, the moment conditions for panel data models considered in this literature are based on the fixed effects approach so they do not deal with the initial conditions problem, which is crucial in the unbalanced case.

The rest of the paper is organized as follows. Section 2 presents the general model and the likelihood functions that account for the unbalancedness. Section 3 formalizes the existing approaches, that is, those ignoring the unbalancedness and making the sample balanced, and discuss the restrictive conditions under which they could work. Section 4 presents the ML and MD estimators for the model that account for the unbalancedness. In Section 5, we study the finite sample properties of the different estimators by means of Monte Carlo simulations. In Section 6, as an empirical illustration, we estimate an export market participation equation using firms' level data. Finally, Section 7 concludes.

II. General framework

We present a general approach that can be applied to dynamic nonlinear panel data models. Let us denote

$$Y_i = (y_{i1},\ldots,y_{iT})',\ X_i = (X'_{i1},\ldots,X'_{iT})',\ S_i = (s_{i1},\ldots,s_{iT})',$$

where $i = 1,\ldots,N$ represents cross-sectional units, y_{it} is the (scalar) outcome, and X_{it} is a row vector of dimension K of covariates. The possibility of having an unbalanced panel is captured through a set of selection indicators, s_{it}:

$$s_{it} = \begin{cases} 1 & \text{if } y_{it} \text{ and } X_{it} \text{ are observed} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the balanced situation can be seen as a particular case of this setting, when $s_{it}=1$ for all i and t. We only consider cases in which either both y_{it} and X_{it} are observed or both are not observed. We define t_i as the first period in which unit i is observed, i.e.

$$t_i=\{t: s_{it}=1 \text{ and } s_{ij}=0\ \forall\, j<t\},$$

and T_i as the number of periods we observe for unit i, $T_i=\sum_{t=1}^{T} s_{it}$. Another characteristic of the panels considered is that all the observations for unit i are consecutive.[3]

Let M_i be the $(T_i \times T)$ matrix that select the set of X_i that we observe, that is, $M_i X_i=(X'_{it_i},\ldots,X'_{iT_i})'$. The element (j,k) of M_i, $m_{i,(jk)}$, is

$$m_{i,(jk)}=\begin{cases} 1 & \text{if } s_{ik}=1 \text{ and } j=k-t_i+1 \\ 0 & \text{otherwise.} \end{cases}$$

If the panel is balanced, M_i is the identity matrix. Note that $S_i=\imath'_{T_i}M_i$ where $\imath_{T_i}$ is a vector of ones with dimension T_i. We denote by J the number of different S_i sequences that we have in the total panel. We refer to the subset of units with the same sequence $S^{(j)}$ as 'sub-panel' $j, j=1,\ldots,J$. Finally, we consider panels where N is large and T and J are small relative to N.

The kind of models we consider in this paper are as follows. For all i and t, y_{it} is determined as

$$y_{it}=g(y_{it-1},X_{it},\eta_i,\varepsilon_{it}),$$

where $g(.)$ is a nonlinear function, non-additively separable in its latent terms,whose form is known up to a vector of parameters that characterized it. For simplicity, we focus on a model with one lag of y_{it} and contemporaneous values of X. However, our analysis could be extended to higher order chains, or to cases that include in X_{it} previous values of these strictly exogenous covariates. 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. Also ε_{it} are 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-constant unobservables, η_i. The function $g(.)$ together with the distribution of ε give the conditional distribution $F(y_{it}\mid y_{it-1},X_i,\eta_i)$ which is our primary object of interest and whose parameters will be estimated. The previous assumptions imply that

$$F(y_{it}\mid y_i^{t-1},X_i,\eta_i)=F(y_{it}\mid y_{it-1},X_{it},\eta_i),$$

where $y_i^{t-1}=(y_{i1},\ldots,y_{it-1})$.

So far this is a standard model in the (balanced) panel data literature. As in Wooldridge (2019), with unbalanced panels, the key assumption that we maintain throughout this paper is

[3] Some other panels present unbalancedness structures that include individuals with non-consecutive observations. In these cases, we could integrate out the holes using the conditional model of y_{it}. But if the model for y_{it} has X covariates that are not observed when $s_{it}=0$, as we consider in this paper, we would need to make further assumptions about X_{it} that would never be made with balanced panels. This is out of the scope of this paper, which is to estimate the model we would specify if having a balanced sample. Nonetheless, if those further assumptions about X are made, the approach in this paper could still be adapted.

$$\varepsilon_{it} \perp S_i \,|\, (\eta_i, X_i) \text{ for all } i \text{ and } t. \tag{1}$$

This implies 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). This means that 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.

Let$f(y_{it} \,|\, y_{it-1}, X_{it}, \eta_i, S_i; \beta)$ be the correctly specified density for the conditional distribution $F(y_{it} \,|\, y_i^{t-1}, X_i, \eta_i)$ under assumption (1), and $h(\eta_i | M_iX_i, S_i; \beta_{\eta S_i})$ the correctly specified density of the distribution $\eta_i \,|\, M_iX_i, S_i$.[4] Then, the density of $(s_{i1}y_{i1}, \ldots, s_{iT}y_{iT})$ for a given individual is

$$\begin{aligned} f(s_{i1}y_{i1}, \ldots, s_{iT}y_{iT} \,|\, M_iX_i, S_i) &= \prod_{t=1}^{T} f(y_{it} \,|\, s_{it-1}y_{it-1}, M_iX_i, S_i)^{s_{it}s_{it-1}} f(y_{it} \,|\, M_iX_i, S_i)^{s_{it}(1-s_{it-1})} \\ &= \left[\prod_{t=t_i+1}^{t_i+T_i-1} f(y_{it} \,|\, y_{it-1}, M_iX_i, S_i) \right] f(y_{it_i} \,|\, M_iX_i, S_i). \end{aligned} \tag{2}$$

Previous equation can be written as

$$\int_{\eta_i} \left[\prod_{t=t_i+1}^{t_i+T_i-1} f(y_{it} \,|\, y_{it-1}, M_iX_i, S_i, \eta_i; \beta) \right] f(y_{it_i} \,|\, M_iX_i, S_i, \eta_i; \lambda_{S_i}) h(\eta_i \,|\, M_iX_i, S_i; \beta_{\eta S_i}) d\eta_i, \tag{3}$$

or as

$$\left[\int_{\eta_i} \prod_{t=t_i+1}^{t_i+T_i-1} f\left(y_{it} \,|\, y_{it-1}, M_iX_i, S_i, \eta_i; \beta\right) h(\eta_i \,|\, y_{it_i}, M_iX_i, S_i; \pi_{\eta S_i}) d\eta_i \right] f\left(y_{it_i} \,|\, M_iX_i, S_i\right), \tag{4}$$

depending on whether we integrate out the unobserved effect by specifying the density for the first observation in each subpanel conditional on the unobserved effect and the density of the unobserved effect, or we specify the density of the unobserved effect conditional on the first observation.

Equations (3) and (4) reveal why it is not trivial to extend the static case considered by Wooldridge (2019) to the dynamic case. With static models, one does not have to deal with the initial conditions problem and the only problem is the potential correlation between the unbalancedness process and the individual effect. In the dynamic case, one has to add the initial conditions problem. This means that$f(y_{it_i} \,|\, M_iX_i, S_i, \eta_i; \lambda_{S_i})$ and $h(\eta_i \,|\, M_iX_i, S_i; \beta_{\eta S_i})$ in equation (3) or $h(\eta_i \,|\, y_{it_i}, M_iX_i, S_i; \pi_{\eta S_i})$ in equation (4) are different for each subpanel. Moreover, writing an equation for$f(y_{i1} \,|\, X_i, \eta_i)$ and $h(\eta_i \,|\, X_i)$, or for $h(\eta_i \,|\, y_{i1}, X_i)$, as Heckman (1981) and Wooldridge (2005) did respectively for the balanced case, is not enough to solve the initial conditions problem for three reasons: (i) the conditioning set of covariates is different for each S_i; (ii) the initial observation is different for each S_i, thus, even in a model without$X$ covariates, without further assumptions$f(y_{it_i} \,|\, \eta_i = \eta, S_i) \neq f(y_{rt_r} \,|\, \eta_r = \eta, S_r)$ for $t_i \neq t_r$; and (iii) even if the starting period t_i is the same, there may be a correlation between S_i and the individual characteristics making the distributions of η_i different for each S_i.

[4] In our notation for defining functions, the set of parameters of that function appear after a semicolon.

III. Existing approaches

Ignoring the unbalancedness

We study under which conditions it is possible to ignore the unbalancedness and to treat the data as if they were balanced. That is, we study when it is possible to use as density of $(s_{i1}y_{i1},\ldots,s_{iT}y_{iT}|M_iX_i)$ the following expression that ignores the unbalancedness,

$$\int_{\eta_i}\left[\prod_{t=t_i+1}^{t_i+T_i-1} f(y_{it}\,|\,y_{it-1},M_iX_i,\eta_i;\beta)\right] f(y_{it_i}\,|\,M_iX_i,\eta_i;\delta)h(\eta_i|M_iX_i;\beta_\eta)d\eta_i, \tag{5}$$

instead of the density given by equation (3) and to have the equivalent Maximum Likelihood Estimators (MLE). We need the following conditions:

1. S_i must be independent of η_i given X, so that $h(\eta_i\,|\,M_iX_i,S_i)=h(\eta_i\,|\,M_iX_i)$ for any set of periods included in X.
2. $h(\eta_i\,|\,M_iX_i)$ must be a function common to all S_i, so that its value changes only as the values of X at which it is evaluated change (but not as a function of the specific periods at which X_i is observed).
3. The process is in the steady state, or the initial observations y_{t_i} come from the same exogenous distribution or rule for all units and t_i.
4. S_i is independent from the shocks to the initial conditions.

Unless these four conditions are all satisfied, the estimates of β obtained by ignoring the unbalancedness are inconsistent. Although very restrictive in general, there are some cases in which condition 1 is satisfied, like rotating panels. Notice that even under this condition 1 $f(y_{it_i}\,|\,X_i,\eta_i,S_i)$ is different for each S_i simply because the process has been running a different number of periods until that first observation, unless we assume that the process is in the steady state. Likewise, $h(\eta_i\,|\,y_{it_i},X_i,S_i)$ will be, in general, different for each t_i. In addition to that, condition 1 is not enough to guarantee that $h(\eta_i\,|\,M_iX_i,S_i)$ is the same for all S_i because, even if S_i is independent of η_i given X, η_i can still depend on X and there will be a different conditioning set of observations in M_iX_i for each S_i.

Condition 2 is very restrictive because, for example, in general $Var(\eta\,|\,M_iX_i)$ will be different if the number periods in which x_{it} are observed is different.[5] A case in which this condition is trivially satisfied is when η_i is independent of X_i.

Conditions 3 and 4 are needed to ensure that all units have the same distribution for the initial condition regardless of the period t_i at which they enter the panel.

Notice that most of these issues arise only in dynamic models, as opposed to what happens in static models as the ones covered in Wooldridge (2019).

Using a subset of periods at which all individuals are observed

As Wooldridge (2005) for dynamic models and Wooldridge (2019) for static models point out, under assumption (1), one could perform the estimation using a subset of data constituting a balanced panel. In particular, he proposes using a subset of periods at which

[5] The condition is not violated, however, if $Var(\eta\,|\,M_iX_i)$ changes with the number of periods of x_{it} observed in a deterministic way, e.g. $Var(\eta\,|\,M_iX_i)=\frac{\sigma_\eta^2}{\iota'_{T_i}M_i\iota_{T_i}}$.

all individuals are observed. Then, one could apply to that balanced sample the standard solutions to the initial conditions problem. Nonetheless, this approach has two limitations: (*i*) it discards useful information leading to an efficiency loss, and (*ii*) the balanced sample may not contain a sufficient number of common periods across individuals, making the estimation infeasible.

Suppose that the correct conditional density of $s_{i1}y_{i1},\dots,s_{iT}y_{iT} \mid M_iX_i,S_i$ is given by (4), excluding the term for the initial observations $f(y_{it_i} \mid M_iX_i,S_i)$. Instead, the following likelihood function is maximized

$$\int_{\eta_i} \left[\prod_{t=t_m}^{T_m} f\left(y_{it} \mid y_{it-1}, X_{it_m}^{T_m}, \eta_i\right)\right] h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, X_{it_m}^{T_m}) d\eta_i, \tag{6}$$

where $t_m \equiv \max_{j\in[1,N]} t_j + 1$, and $T_m \equiv \min_{j\in[1,N]}(t_j + T_j - 1)$. Under Assumption 1 $f\left(y_{it} | y_{it-1}, X_{it_m}^{T_m}, S_i, \eta_i\right) = f\left(y_{it} \mid y_{it-1}, X_{it_m}^{T_m}, \eta_i\right)$. Thus, to have a consistent ML Estimator of the parameters of the conditional distribution of $y_{it} \mid y_{it-1}, M_iX_i, \eta_i$ based on (6), we need

$$\begin{aligned} & h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, X_{it_m}^{T_m}) \\ & = \sum_{j=1}^{J} h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, X_{it_m}^{T_m}, S_i = S^{(j)}) \Pr\left(S_i = S^{(j)} \mid y_{i\max_{j\in[1,N]} t_j}, X_{it_m}^{T_m}\right), \end{aligned} \tag{7}$$

where $S^{(j)}$ is the j-th element of the set of J different S_i sequences that we have in the panel, and $X_{it_m}^{T_m} = \left(X'_{it_m},\dots,X'_{i,T_m}\right)'$. So, as long as the $h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, X_{it_m}^{T_m})$ we specify satisfies this condition and we have enough periods in the balanced sample, the MLE based on (6) will be consistent, though less efficient. However, depending on the nature of $h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, M_iX_i, S_i)$ (i.e. depending on the nature of the relation between η_i and S_i and the evolution of the distribution of y_{it} across periods and subpanels) approximating $h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, X_{i\max t_i+1}^{\min(t_i+T_i-1)})$ may require a complex distribution even if $h(\eta_i \mid y_{i\max_{j\in[1,N]} t_j}, M_iX_i, S_i)$ is the standard normal distribution.

Using a subset of individuals observed the same periods

Another possibility to deal with the unbalancedness, found in the applied literature using both static and dynamic models, is to take one single subpanel from the total sample. In many cases, this would be the subsample of individuals present in all the waves of the original panel (as in Contoyannis *et al.*, 2004). More generally, one can take the subset of individuals observed only in some specific consecutive waves.

Although this way of obtaining a balanced sample produces an efficiency loss because it discards a potentially high proportion of the sample, it avoids the infeasibility of the previous balancing method and may consistently estimate the common parameters of the model. However, the average marginal effects we estimate for this subsample are not a consistent estimation of the average marginal effects for the entire sample. The reason is that the conditional distribution of the heterogeneous individual effects will only be valid for this particular subgroup of individuals, unless the unbalancedness is independent of η_i and of X. The distribution of η_i for this balanced subsample is different from the distribution of η_i for the entire sample. And the marginal effects, which are the ultimate parameters of interest, are a function of the distribution of η_i.

IV. Estimation

Maximum likelihood estimation

The models that account for unbalancedness explained in Section 2 can be estimated by Maximum Likelihood (ML). The log-likelihood, if the model specifies the terms in (3), is given by

$$\mathcal{L}=\sum_{i=1}^{N}\log\int_{\eta_i}\left[\prod_{t=t_i+1}^{t_i+T_i-1} f\big(y_{it}|y_{it-1},M_iX_i,S_i,\eta_i;\beta\big)\, f\big(y_{it_i}|M_iX_i,S_i,\eta_i;\lambda_{S_i}\big)\, h(\eta_i|M_iX_i,S_i;\beta_{\eta S_i}\big)\right] d\eta_i. \tag{8}$$

If the model instead specifies the terms in (4), the log-likelihood is given by

$$\mathcal{L}=\sum_{i=1}^{N}\log\int_{\eta_i}\left[\prod_{t=t_i+1}^{t_i+T_i-1} f\left(y_{it}\,|\,y_{it-1},M_iX_i,S_i,\eta_i;\beta\right) h(\eta_i\,|\,y_{it_i},M_iX_i,S_i;\pi_{\eta S_i})\right] d\eta_i. \tag{9}$$

These log-likelihood functions will be maximized with respect to the vector of parameters $\theta=\left(\beta',\gamma'\right)'$ that can be partitioned into the set of common parameters β and the set of subpanel specific parameters $\gamma=\left(\gamma_1',\ldots,\gamma_J'\right)'$. The specific parameters to subpanel j are $\gamma_j=(\lambda'_{S^{(j)}},\beta'_{\eta S^{(j)}})'$ in (8), or $\gamma_j=\pi_{\eta S^{(j)}}$ in (9).

The properties of the MLE are well known, as well as the numerical procedures to obtain it. The problem is that the optimization procedure is cumbersome. Our specific likelihood must be optimized jointly with respect to a high number of parameters, because, due to the unbalancedness, there is a different set of some parameters for each subpanel. This will typically preclude using standard estimation software and will increase the computation time.

Minimum distance estimation

We propose an estimation method that keeps the good asymptotic properties of the MLE but reduces its computational burden. Moreover, it allows us to use the same routine or estimation programme as when having a balanced panel.

The proposal has two steps. The first step is to estimate the model for each subpanel separately. This implies that we can use the same standard software as in balanced panels and accommodate very easily different distributions of η_i for each subpanel S_i.

The second step is to obtain estimates of the parameters $\theta=(\beta',\gamma')'$ by Minimum Distance (MD). Let $\hat{\delta}=(\hat{\delta}_1',\hat{\delta}_2',\ldots,\hat{\delta}_J')'$ be the vector of estimated coefficients of the model after the first step. Each $\hat{\delta}_j'$ includes two types of parameters: $\hat{\delta}_j^{[c]}$, the estimates of the parameters β that are common across subpanels, and $\hat{\delta}_j^{[nc]}$, the estimates of the non-common parameters γ_j.

In order to recover a unique estimate of β, we set the vector δ to be equal to a known function of the structural parameters θ : $\delta=h(\theta)$ where $h(.)$ restricts all the $\hat{\delta}_j^{[c]}$ to be estimates of the same β parameters. The structural parameters θ can be consistently and efficiently estimated by minimizing the following quadratic form:

$$\hat{\theta}^{MD} = \arg\min_{\theta} Q(\theta) = \left[\hat{\delta} - h(\theta)\right]' V^{-1} \left[\hat{\delta} - h(\theta)\right], \tag{10}$$

where V is the var-cov matrix of $\hat{\delta}$, which is a block diagonal matrix since different subpanels have no observations in common. See Appendix S1 for further details on the MD estimator and its algebraic expressions for our case.

This procedure is known to be asymptotically equivalent to maximizing the log likelihood $\mathcal{L}$ on the entire set of parameters θ (see Chamberlain, 1982 and 1984 and references there in). If $N \rightarrow \infty$ but T and J are fixed, then the asymptotic properties derived in those references are applicable to our case. These are the relevant conditions for us since we are interested in situations in which N is large relative to T and J. Then $\hat{\theta}^{MD}$ is asymptotically equivalent to $\hat{\theta}^{MLE}$.

Average marginal effects

Once we have estimated the parameters of the model and of the distribution of the unobserved heterogeneity, either by ML or by MD, we use them to obtain estimates of Average Marginal Effects (AMEs), which are the ultimate parameters of interest. The crucial aspect of this paper is that the average and the distribution of the unobserved heterogeneity used to estimate the AMEs are conditional on the unbalancedness structure that we have. Failing to account for it and, especially, for the potential correlation between the unbalancedness and the individual effect, will result in biases in the estimates of the AMEs as well.

In Albarran, Carrasco and Carro (2018), we show how to implement the estimators presented in previous sections, as well as the AMEs, with specific assumptions about parametric distributions.

V. Simulations: Finite sample performance

We use Monte Carlo techniques to study the finite sample performance of the estimators under different degrees of unbalancedness. In this section, we consider a dynamic model without other covariates. The next section presents simulation results with exogenous covariates based on the data used in the empirical application.

The subpanels may vary in both the period the individuals enter and when they leave the sample. The degree of unbalancedness in the sample is governed by J, which, as defined in Section 2, indicates the number of subpanels. In our baseline DGP, $J=2$ indicates that there are two subpanels: the first half of units $(\frac{N}{2})$ are observed from 1 to $T-1$ and the second half of units are observed from 2 to T. When $J=4$ the first quarter of units are observed from 1 to $T-1$, the second from 1 to $T-2$, the third from 2 to T, and the last quarter is observed from 3 to T. The same for higher values of J. Given this way of generating the unbalancedness, J can only take even values. We impose the following restrictions on the values of J: (i) the maximum value is $J_{\max} = \min\{2 * T - 3, \frac{N}{30}\}$, where

$2*T-3$ guarantees that all subpanels have at least 3 periods and $\frac{N}{30}$ guarantees that there is at least 30 units in all subpanels, and (*ii*) the minimum value is $J_{\min}=\max\{2*T-15, 0\}$, where the restriction $2*T-15$ is to have at least one subpanel with less than 8 periods.[6]

Our baseline Data Generating Process (DGP) is as follows:

$$y_{it}=1\{\alpha y_{it-1}+\eta_i+\varepsilon_{it}\geqslant 0\}, \tag{11}$$

where ε_{it} and η_i follow normal distributions of the form:

$$\varepsilon_{it}\underset{iid}{\sim}N(0,1),\ \eta_i \,|\, S_i\in j\underset{iid}{\sim}N(\mu_{\eta j},\sigma^2_{\eta j}), \tag{12}$$

$$\mu_{\eta j}=\mu_\eta+\big(1.3*J/(J-1)\big)*\big((j/J)-(J+1)/(2*J)\big), \tag{13}$$

$$\sigma_{\eta j}=0.25+(j-1)*\ \big((\sigma_\eta-0.2)/(J-1)\big), \tag{14}$$

$$y_{i0}=1\{\pi_0+\big((j/J)-\big(J+1\big)/\big(2*J\big)\big)+\pi_1\eta_i+v_{i0}\geqslant 0\}, v_{i0}\underset{iid}{\sim}N(0,1), \tag{15}$$

where $\alpha=0.75$, $\mu_\eta=0$, $\sigma^2_\eta=1$, $\pi_0=-1.25$, and $\pi_1=0.5$, so the initial condition and η_i are both correlated with the unbalancedness process. Moreover, $E_j(\mu_{\eta j})=\mu_\eta=0$, $E_j(\sigma_{\eta j})=0.6$, $\mu_{\eta j}\in[-1,1]$, $\sigma_{\eta j}\in[0.2,1]$, and $\mu_{\eta j}$ and $\sigma_{\eta j}$ are increasing in S, so that the value of η_i is more likely to be greater the greater the value of j, i.e. for the last subpanels.[7] We consider unbalancedness to the right and to the left (that is, subpanels differ in both the period individuals enter and leave the sample) and we have run 1,000 replications for each DGP with $N=1,000$.

For the sake of brevity not all estimators are used in all the simulation exercises. Except for the highly used solution of selecting the longest subpanel, our general criteria has been to avoid simulating and/or reporting estimators that are not correct given the assumptions made in each DGP. Nevertheless, Table 5 reports the incorrect estimator that ignores the unbalancedness too, so the problem of ignoring the unbalancedness can be evaluated in a realistic context.

Tables 1 and 2 report means and root mean squared errors (RMSE) for the α parameter and the AME, respectively. Since the true AME (slightly) varies with the sample drawn in each Monte Carlo simulation, Table 2 also reports the true expected AME. We deal with the initial conditions problem by specifying the density of the unobserved effect conditional on the first observation.

We report results for the ML estimator making the panel balanced using the subset of periods at which all individuals are observed (labelled 'Bal. Periods'), for the ML estimator making the panel balanced using the subset of individuals that are observed in the same period (labelled 'Bal. Units'), and for the MD and ML estimators that account for the unbalancedness and for its correlation with η_i (labelled as 'Unbal. MD' and 'Unbal. ML', respectively).

[6] When the time length is long, the fixed effects approaches may be preferable. For example, simulations in Carro (2007) show cases where a modified MLE fixed effects estimator performs well with 8 periods.

[7] Further details about the DGP and on how the unbalancedness structure of the data has been generated can be found in Appendix S2.

TABLE 1

Simulation results on the estimation of α. Baseline case. Double Unbal.

	Bal. Periods	*Bal. Units*	*Unbal. MD*	*Unbal. ML*	*FE, Ignore Unbal.*	*SPJ, Bal. Units*	*Bal. Periods*	*Bal. Units*	*Unbal. MD*	*Unbal. ML*	*FE, Ignore Unbal.*	*SPJ, Bal. Units*
				$\hat{\alpha}$					*RMSE*			
T=4												
J=2		0.7484	0.7361	0.7185				0.1844	0.1241	0.1283		
T=6												
J=2	0.7211	0.7505	0.7492	0.7468	−0.1755		0.0925	0.1021	0.0685	0.0675	0.9279	
J=6		0.7488	0.7505	0.7356	−0.5119			0.1827	0.0811	0.0810	1.2656	
T=8												
J=4	0.7457	0.7457	0.7512	0.7499	0.1368	1.0522	0.0925	0.1146	0.0506	0.0504	0.6153	0.3491
J=10		0.7454	0.7512	0.7407	−0.0758	1.0917		0.1804	0.0638	0.0638	0.8287	0.5362
T=10												
J=6	0.7502	0.7433	0.7515	0.7507	0.2795	0.8682	0.1004	0.1210	0.0430	0.0433	0.4724	0.1988
J=10		0.7461	0.7533	0.7505	0.2063	0.8711		0.1614	0.0494	0.0489	0.5458	0.2366
J=14		0.7430	0.7516	0.7430	0.1204	0.8729		0.1929	0.0549	0.0544	0.6321	0.2867
T=15												
J=16		0.7458	0.7492	0.7484	0.4080	0.8247		0.1535	0.0353	0.0352	0.3437	0.1952

TABLE 2

Simulation results on the estimation of the AMEs. Baseline case. Double Unbal.

		Bal. Periods	*Bal. Units*	*Unbal. MD*	*Unbal. ML*	*FE, Ignore Unbal.*	*SPJ, Bal. Units*	*Bal. Periods*	*Bal. Units*	*Unbal. MD*	*Unbal. ML*	*FE, Ignore Unbal.*	*SPJ, Bal. Units*
	AME	$\widehat{AME}$						*RMSE*					
T=4													
J=2	0.2194		0.1804	0.2201	0.2072				0.0634	0.0460	0.0441		
T=6													
J=2	0.2194	0.2152	0.1810	0.2259	0.2183	−0.0591		0.0315	0.0481	0.0291	0.0228	0.2795	
J=6	0.2292		0.1833	0.2320	0.2230	−0.1733			0.0696	0.0312	0.0289	0.4036	
T=8													
J=4	0.2280	0.2414	0.1820	0.2349	0.2279	0.0457	0.3421	0.0381	0.0567	0.0205	0.0176	0.1832	0.1270
J=10	0.2300		0.1833	0.2329	0.2253	−0.0254	0.3425		0.0698	0.0240	0.0224	0.2565	0.1425
T=10													
J=6	0.2300	0.2484	0.1815	0.2368	0.2300	0.0923	0.2715		0.0596	0.0174	0.0149	0.1384	0.0659
J=10	0.2304		0.1834	0.2362	0.2303	0.0686	0.2705		0.0664	0.0190	0.0171	0.1628	0.0758
J=14	0.2303		0.1831	0.2342	0.2270	0.0403	0.2699		0.0725	0.0209	0.0190	0.1910	0.0885
T=15													
J=16	0.2312		0.1844	0.2372	0.2304	0.1334	0.2376		0.0653	0.0137	0.0118	0.0986	0.0568

In Table 1, we observe that the four approaches considered provide estimated values of the parameter α very close to its true value. However, there exists some relevant points that are worth noting. The 'Bal. Periods' solution has two important drawbacks compared to the approaches that account for the unbalancedness. First, it cannot be used in many cases, including some for which the unbalancedness is moderate. Second, it implies an important loss of efficiency when it can be employed, even for moderate unbalancedness. Regarding the 'Bal. Units' estimator, the RMSE is much higher than in the case of any other estimator due to the loss of observations when using this estimator. Furthermore, in the estimation of the AME, it not only presents an efficiency loss but also a bias problem: the RMSE is twice to five times larger than the 'Unbal. MD' and the 'Unbal. ML' .

The comparison between 'Unbal. MD' and 'Unbal. ML' shows that, as expected, their behaviour is very similar. Nonetheless, estimating the model by ML is computationally cumbersome: it takes between 150 and 1,600 times more computing time than the MD, depending on the number of periods and subpanels. On the other hand, we can face a potential problem of lack of variability in certain subpanels when estimating by MD. In our simulations, the percentage of failures due to lack of variability is below 1%. Higher failure rates only appear in a few cases when considering a very high degree of state dependence.

For completeness, Tables 1 and 2 report simulations using a Fixed Effects estimator without bias correction, and using the Split Panel Jackknife (SPJ) proposed by Dhaene and Jochmans (2015) and also used in Fernandez-Val and Weidner (2016) to correct the bias. Even with the bias correction, the bias is very large when compared with the MD estimator, probably due to many subpanels having a small number of periods. Taking only the longest subpanel to try to reduce the incidental parameters problem introduces a sample selection bias in the estimation of the AMEs. This leads to a bias and RMSE on the SPJ estimates of the AME three to six times higher than those of the MD. This reinforces the CRE as an useful alternative to estimate the models considered in this paper with unbalanced panels.

Appendix S3 presents a number of simulation results where we have sequentially changed different parameters of the baseline DGP. In subsection C.I we have simulated the baseline DGP with left-side unbalancedness (i.e., subpanels differ only in the period they start but all are observed until T). We have reduced the sample size to $N=500$ in subsection C.II and changed the degree of state dependence in subsection C.III. Finally subsection C.IV presents results in which the initial condition and/or the unbalancedness are uncorrelated with η_i and estimators that deal with the initial conditions problem by specifying the density of the first observation conditional on η_i and the density of η_i. Independence of the unbalancedness from η_i leads to some simplification in the estimators when taken into account, although the estimator presented in section 4.2 is still valid because it does not assume anything about the relation between S_i and η_i. Apart from the RMSE of all estimators increasing as the sample size decreases, the results presented so far remain basically unchanged.

VI. An application to export market participation

We illustrate previous methods by estimating a model for firms' export market participation decision. We use data for Spanish manufacturing firms, the Business Strategies Survey

(*Encuesta sobre Estrategias Empresariales*, ESEE) for the period 1990–1999.

Our sample consists of an unbalanced panel with 14 different subpanels of 1,807 firms and 12,683 observations.[8] The starting point for estimation is an equation of the form

$$y_{it} = \mathbf{1}(\alpha y_{it-1} + X'_{it}\beta + \eta_i + v_{it} \geqslant 0), \tag{16}$$

where $y_{it} = 1$ if the $i - th$ firm exported in year t. Our empirical model is based on a simple model of optimization for a firm facing the export decision (see Roberts and Tybout, 1997). The choice of variables included in the vector X largely follows the previous literature on the determinants of firm's export decisions.

Tables 3 and 4 present the AMEs for the lagged export status variable.[9] With this data set, we cannot perform the estimates using the 'Bal. Periods' solution due to lack of observations. Column labelled 'Bal. Units' presents the estimates using the balanced sample in which firms are observed all periods (subpanel $S = 8$). Column labelled 'Ignore Unbal.' presents the estimates that ignore the unbalancedness. The last column presents the results from the model that accounts for the unbalancedness and allows for its correlation with the unobserved effect. We model the unobserved heterogeneity conditional on the initial condition and the time average of the exogenous variables.

First row in Table 3 presents the AME's for the entire sample. Coincidentally the 'Ignore Unbal.' and the 'Bal. Units' estimators provide similar results, probably because the observations in the balanced sample are 56% of the sample used to estimate the model ignoring unbalancedness. This is an example where comparing these two estimators might lead to the incorrect conclusion about the possibility of ignoring the unbalancedness.[10]

Regarding the 'Unbal. MD' estimator, we find that the estimated AME for the entire sample is around 4 percentage points greater than the one from the 'Ignore Unbal.' estimator. This difference is statistically significant even though the AME for the entire sample tends to mask biases in opposite directions in different subsamples.[11] This can be seen in more detail if we analyse the AMEs by subgroups (see Table 3) and by subpanels (see Table 4). In particular, we find statistically different results for younger firms and for firms that do not export in the first period. Last row in Table 3 presents the estimates excluding the largest subpanel ($S = 8$) to show that if we had a data set without a subpanel that dominates so much, the differences between 'Ignore Unbal.' and 'Unbal. MD.' are more significant.

If we look at the AMEs by subpanel, there are five in which the MD gives statistically significant AMEs and the differences are even larger than for the total sample. Furthermore, while the MD estimates range between 0.1095 and 0.4399, the corresponding estimates for the model ignoring the unbalancedness only range from 0.2108 to 0.2689. Thus, there

[8] See Appendix S4 for further details, including definition and descriptive statistics of variables used in the application.

[9] Table D.3 in Appendix S4 presents the estimates of the common parameters of the model.

[10] If normality about the distribution of η_i is incorrectly assumed in both cases, these two estimators will tend to produce similar biased estimates. Therefore, the comparison between them may lead to the incorrect conclusion that the unbalancedness can be ignored (see Albarran, Carrasco and Carro, 2018).

[11] Last column of Tables 3 and 4 presents a Hausman-type test of the difference between the 'Unbal. MD' and the 'Ignore Unbal.' estimators using the variance–covariance matrix of the MD estimates only instead of subtracting from it the variance of the 'Ignore Unbal.' estimator. Under correct specification, this represents a lower bound for this test and a rejection here will also be a rejection when using the well-defined variance–covariance matrix of the difference.

TABLE 3

Estimated Average marginal effects of Lagged Export.

	Bal. Units *(1)*	*Ignore Unbal.* *(2)*	*Unbal. MD* *(3)*	*Test of Diff.* *(2) vs (3)*
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)	

Note: Standard errors are reported in parentheses. The implementation of the test of difference is discussed in footnote 6. Asterisks indicate the difference is significantly different from zero at *10%; **5%; ***1%.‡ Approximately one third of the firms are younger than 12 and around 40% are 24 or older.

is a great deal of variation on the marginal effect of lagged export across subpanels that is not captured by the 'Ignore Unbal'. estimator. These results indicate that the model that ignores the unbalancedness incorrectly imposes, among other restrictions, independence between the distribution of the unobserved heterogeneity and the unbalancedness.

Simulation evidence on the properties of the estimators We simulated data calibrated to the ESEE sample to study the properties of the estimators in the empirical application. This also has the additional interest of exhibiting some Monte Carlo results with covariates in the dynamic model, which complement those reported in Section 5.

The DGP is extended here to incorporate exogenous covariates. Thus, the main equation becomes

$$y_{it} = 1\{\alpha y_{it-1} + X_{it}'\beta + \eta_i + \varepsilon_{it} \geqslant 0\}, t = t_i + 1, \ldots, t_i + T_i, \tag{17}$$

$$\varepsilon_{it} \mid y_{it_i}, X_i, S_i \in j \underset{iid}{\sim} N(0,1),\ \eta_i \mid y_{it_i}, X_i, S_i \in j \sim N\left(\pi_{0j} + \pi_{1j} y_{it_i} + \overline{\overline{X_i}}' \pi_{2j}, \sigma^2_{\eta j}\right), \tag{18}$$

where X_{it} denotes the vector of exogenous regressors, $\overline{\overline{X_i}}$ contains the within-means (from period $t_i + 1$ to $t_i + T_i$) of the time-varying explanatory regressors and y_{it_i} is the first observed value of the endogenous variable for the individual i.

Table 5 contains the simulation results. The results confirm that the 'Unbal. MD' outperforms both 'Bal. Units' and 'Ignore Unbal'. estimators. In part B of Table 3, we check the extent to which each estimator is able to capture heterogeneity in the AME across subgroups. As expected, the 'Bal. Units' estimator does a nice work in the only subpanel

TABLE 4

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	
	(0.0290)	(0.0233)	(0.0245)	
$S=9$		0.2195	0.2758	
		(0.0221)	(0.0793)	
$S=10$		0.2612	0.2634	
		(0.0257)	(0.1403)	
$S=11$		0.2689	0.3256	
		(0.0260)	(0.0830)	
$S=12$		0.2674	0.3144	
		(0.0251)	(0.1175)	
$S=13$		0.2563	0.4399	***
		(0.0250)	(0.0393)	
$S=14$		0.2374	0.3765	
		(0.0239)	(0.0877)	

Note: See note in Table 3.

that uses, but it neglects the other ones. On the other hand the 'Ignore Unbal'. estimator is able to provide different AMEs across subgroups, but they are substantially biased in some of them. By contrast, the 'Unbal. MD' estimator performs reasonably well overall.

VII. Conclusion

In this paper, we consider the estimation of dynamic nonlinear CRE models when using unbalanced panel data. We identify two types of problems: (i) an inconsistency in the estimates of the coefficients when the unbalancedness is ignored; and (ii) an efficiency loss and/or an inconsistency in the estimates when using different balanced versions of the unbalanced original data. These problems are especially severe when the unbalanced process is correlated with the individual effect.

TABLE 5

Simulation results based on the MD results obtained in the empirical application

True parameter	*Bal. Units*		*Ignore Unbal.*		*Unbal. MD*	
	Estimated	*RMSE*	*Estimated*	*RMSE*	*Estimated*	*RMSE*
A. Total Sample						
$\alpha = 1.5153$	1.5145	0.0962	1.4651	0.0900	1.5413	0.0796
$AME = 0.2721$	0.2332	0.0483	0.2449	0.0358	0.2539	0.0305

	True AME	*Bal. Units*		*Ignore Unbal.*		*Unbal. MD*	
		$\widehat{AME}$	RMSE	$\widehat{AME}$	RMSE	$\widehat{AME}$	RMSE
B. By Subgroups							
By Age							
<12	0.3055	0.2495	0.0639	0.2606	0.0513	0.2698	0.0463
$12-24$	0.2895	0.2606	0.0425	0.2647	0.0352	0.2761	0.0292
>24	0.2310	0.2065	0.0363	0.2172	0.0258	0.2249	0.0245
By Initial Conditions							
$y_{t_i} = 1$	0.2167	0.1644	0.0580	0.1974	0.0287	0.2041	0.0281
$y_{t_i} = 0$	0.3200	0.2962	0.0412	0.2857	0.0430	0.2970	0.0364
By Subpanels							
$S = 8$	0.2334	0.2332	0.0280	0.2442	0.0255	0.2387	0.0266
$S \neq 8$	0.3196			0.2456	0.0778	0.2726	0.0576

We propose a general model that accounts for the unbalancedness that can be arbitrarily correlated with the permanent unobserved heterogeneity. We show that this model can be estimated by ML and also by MD. 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 favour the MD approach.

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 nonlinear model.

Final Manuscript Received: May 2018

References

Akee, R. K., Copeland, W. E., Keeler, G., Angold, A. and Costello, E. J. (2010). 'Parents' incomes and children's outcomes: a quasi-experiment using transfer payments from casino profits', *American Economic Journal: Applied Economics*, Vol. 2, pp. 86–115.

Albarran, P., Carrasco, R. and Carro, J. (2018). Using Stata to Estimate Dynamic Nonlinear Random Effects Models with Unbalanced Panels, mimeo.

Bhattacharya, D. (2008). 'Inference in panel data models under attrition caused by unobservables', *Journal of Econometrics*, Vol. 144, pp. 430–446.

Carro, J. M. (2007). ‘Estimating dynamic panel data discrete choice models with fixed effects’, *Journal of Econometrics*, Vol. 140, pp. 503–528.

Chamberlain, G. (1982). ‘Multivariate regression models for panel data’, *Journal of Econometrics*, Vol. 18, pp. 5–46.

Chamberlain, G. (1984). ‘Panel data’, in Griliches Z., Intrilligator M. D., (eds.), *Handbook of Econometrics*, Vol. 2. North-Holland: Amsterdam.

Contoyannis, P., Jones, A. M. and Rice, N. (2004). ‘The dynamics of health in the british household panel survey’, *Journal of Applied Econometrics*, Vol. 19, pp. 473–503.

Dhaene, G. and Jochmans, K. (2015). ‘Split-panel jackknife estimation of fixed-effect models’, *The Review of Economic Studies*, Vol. 82, pp. 991–1030.

Fariñas, J. C. and Jaumandreu, J. (1999). ‘Diez años de encuesta sobre estrategias empresariales’, *Economía Industrial*, Vol. 329, pp. 29-42.

Fernandez-Val, I., and Weidner, M. (2016). ‘Individual and time effects in nonlinear panel data models with large N,T’, *Journal of Econometrics*, Vol. 192, pp. 291–312.

Heckman, J. J. (1981). ‘The incidental parameters problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process’, in Manski, C. and McFadden, D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA: MIT Press, pp. 114–178.

Hyslop, D. R. (1999). ‘State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women’, *Econometrica*, Vol. 67, pp. 1255–1294.

Pacini, D. and Windmeijer, F. (2015). Moment conditions for AR(1) panel data models with missing outcomes, Discussion Paper 15/660, Department of Economics, University of Bristol.

Roberts M., and Tybout, J. (1997). ‘The decision to export in Colombia: An empirical model of entry with Sunk Costs’, *American Economic Review*, Vol. 87, pp. 545–564.

Stewart, M. B. (2007). ‘The interrelated dynamics of unemployment and low-wage employment’, *Journal of Applied Econometrics*, Vol. 22, pp. 511–531.

Wooldridge, J. M. (2005). ‘Simple solutions to the initial conditions problem for dynamic, nonlinear panel data models with unobserved heterogeneity’, *Journal of Applied Econometrics*, Vol. 20, pp. 39–54.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems’, *Journal of Econometrics*, Vol. 141, pp. 1281–1301.

Wooldridge, J. M. (2019) ‘Correlated random effects models with unbalanced panels’, *Journal of Econometrics*, Vol. 211, pp. 137–150.

Supporting Information

Additional supporting information may be found in the online version of this article:

Supplementary Appendices with further details and results.