Journal of Econometrics 211 (2019) 137–150

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Correlated random effects models with unbalanced panels

Jeffrey M. Wooldridge[1]

Department of Economics, Michigan State University, East Lansing, MI 48824-1038, United States

ARTICLE INFO

Article history:
Available online 10 December 2018

JEL classification:
C13
C23

Keywords:
Correlated random effects
Panel data
Unbalanced panel
Hausman test

ABSTRACT

I propose some strategies for allowing unobserved heterogeneity to be correlated with observed covariates and sample selection for unbalanced panels. The methods are extensions of the Chamberlain–Mundlak approach for balanced panels when explanatory variables are strictly exogenous conditional on unobserved effects. A byproduct is fully robust Hausman tests for unbalanced panels. Even for nonlinear models, in many cases the estimators can be implemented using standard software. The framework suggests straightforward tests for sample selection that is correlated with unobserved shocks while allowing selection to be correlated with the observed covariates and unobserved heterogeneity.

1. Introduction

Correlated random effects (CRE) approaches to nonlinear panel data models are popular with empirical researchers, partly because of their simplicity but also because recent research [for example, Altonji and Matzkin (2005) and Wooldridge (2005)] shows that quantities of interest – usually called "average partial effects" (APEs) or "average marginal effects" (AMEs) – are identified under nonparametric restrictions on the distribution of heterogeneity given the covariate process. (Exchangeability is one such restriction, but it is not the only one.) Wooldridge (2010) shows how the CRE approach applies to commonly used models, such as unobserved effects probit, Tobit, and count models. Papke and Wooldridge (2008) propose simple CRE methods when the response variable is a fraction or proportion.

Leading competitors to CRE approaches are so-called "fixed effects" (FE) methods, which treat the heterogeneity as parameters to be estimated. As is well known, except in some special cases, estimating unobserved heterogeneity for each unit in the sample generally suffers from the incidental parameters problem. Headway has been made in obtaining bias-corrected versions of fixed effects estimators for nonlinear models — for example, Hahn and Newey (2004) and Fernández-Val (2009). These methods are promising, but they currently have some practical shortcomings. First, the number of time periods needed for the bias adjustments to work well may be greater than is available in many applications. (Technically, the adjustments are obtained assuming $T \to \infty$.) Second, many of the bias corrections assume stationarity in the time series dimension, and they all require weak dependence; in some cases, the untenable assumption of serial independence is maintained. Stationarity rules out common staples in empirical work, such as including dummy variables to capture secular time effects. Weak dependence is maintained because the asymptotic analysis relies on both T and N getting large.

More recently, Chernozhukov et al. (2013) (CFHN) show that average partial effects are not generally identified in nonlinear models, and they provide estimable bounds in the case of discrete covariates. Under stationarity and ergodicity,

E-mail address: wooldri1@msu.edu.

[1] Paper presented at the Conference in Honor of Jerry Hausman, Chatham, MA, October 1–3, 2015. I would like to thank two anonymous referees, Whitney Newey, and the conference participants for helpful comments on the previous draft. Simon Quinn and Stefanie Schurer provided useful suggestions on the first draft, which was presented at the 15th International Panel Data Meetings in Bonn, Germany, 2009.

https://doi.org/10.1016/j.jeconom.2018.12.010
0304-4076/

CFHN show that the bounds become tighter as the number of time periods (T) increases. [They do not impose assumptions such as exchangeability, as in Altonji and Matzkin (2005), in which case the APEs are point identified.] These methods are very promising but are still limited to discrete covariates. Plus, in some cases we may be willing to impose more restrictions in order to point identify the APEs for any number of time periods.

Currently, the most general framework for bias-adjusted FE estimators is Fernández-Val and Weidner (2016), which allows for both time and cross-sectional fixed effects, does not require independence over time, and applies to models under sequential exogeneity. Nevertheless, weak dependence is imposed, and the methods are computationally difficult. I view the CRE approach as complementary to fixed effects/bias adjustment approaches, with CRE applying in situations with short panels, arbitrary time heterogeneity, and arbitrary time dependence. In convincing empirical applications, several modeling and estimation strategies are applied, and it is especially useful to compare methods that are justified under nonnested sets of assumptions.

Another competitor to CRE approaches, but only in special cases, is conditional maximum likelihood estimation. When it applies, the CMLE has the advantage that it puts no restrictions on the heterogeneity distribution — either unconditionally or conditionally. Further, it works for any number of time periods and imposes no restrictions on the time series properties of the covariates. However, because CMLEs leave heterogeneity distributions unspecified, it is unclear how to obtain average partial effects, especially for small T. More importantly, even in the limited cases where it applies, the CMLE imposes conditional independence across time, and the assumption appears to be critical in the leading case of the unobserved effects logit model. Kwak et al. (2018) show that the CMLE for the unobserved effects logit model is severely biased if the conditional independence assumption fails.

In the balanced panel case, CRE approaches put restrictions on the conditional distribution of heterogeneity given the entire history of the covariates. This is its drawback compared with FE or CMLE approaches. But when explanatory variables are strictly exogenous, it requires few other assumptions for estimating average partial effects, and the restrictions needed on the conditional heterogeneity distribution can be fairly weak. For example, stationarity and weak dependence of the processes over time are not necessary. In other words, for estimation using balanced panels, CRE, FE, and CMLE involve tradeoffs among assumptions and the types of quantities that can be estimated. No method provides consistent estimators of either parameters or APEs under a set of assumptions strictly weaker than the assumptions needed for the other procedures.

One clear disadvantage of CRE approach compared with the FE and CMLE methods is that the latter approaches do not require modification for unbalanced panels, whereas CRE methods, as currently developed for nonlinear models, are for balanced panels. Generally, it is not obvious how one should extend the CRE approach to account for unbalanced panels. The purpose of the current paper is to propose an approach that can be applied to commonly used parametric models. It will be clear that the approach can be made less parametric, but that is not formally pursued in this paper.

A key assumption used in this paper is that sample selection is not systematically related to unobserved shocks. This assumption is either implicit or explicit in most analyses with unbalanced panels, particularly when heterogeneity is removed or treated as parameters to estimate. Nevertheless, one of the attractions of, say, fixed effects estimation in the linear model – which we study from the CRE perspective in Section 2 – is that selection can be arbitrarily correlated with unobserved heterogeneity. My general approach to CRE models also allows such correlation. In fact, the heterogeneity is allowed to be correlated with the entire history of the selection indicators and the selected covariates. Unlike CMLE approaches, I do not restrict the serial dependence in the data.

This paper is not intended to offer missing data corrections when selection is correlated with the idiosyncratic errors. Doing so is clearly worthwhile, and has been done in some cases where exogenous variables are always observed [for example, Wooldridge (1995) and Semykina and Wooldridge (2010) for the linear model, Semykina and Wooldridge (2017) for a probit model]. Obtaining valid corrections when data can be missing on covariates is much more challenging. This paper provides a starting point by providing a systematic treatment of missing data in nonlinear panel data models where data may be missing on both the explanatory and response variables, and selection is correlated with unobserved heterogeneity.

One important restriction in this paper is that, conditional on unobserved heterogeneity, the explanatory variables are strictly exogenous. In the linear case, Abrevaya (2013) shows how the Chamberlain (1982) projection approach can be extended to unbalanced panels in cases where the explanatory variables are either strictly or sequentially exogenous. Abrevaya's approach can be viewed as an alternative to the Arellano and Bond (1991) approach to linear models with a single, additive source of heterogeneity. The current paper allows for many sources of heterogeneity and shows how the CRE approach can be applied to nonlinear models.

The rest of the paper is organized as follows. Section 2 studies the behavior of estimators for unbalanced panels for the standard linear model with an additive unobserved effect. Somewhat surprisingly, adding the time average of the covariates (averaged across the unbalanced panel) and applying either pooled OLS or random effects still leads to the fixed effects (within) estimator, even when common coefficients are imposed on the time average. One reason the algebraic equivalence is useful is because it generates simple, fully robust Hausman specification tests for choosing between the random effects (RE) and fixed effects (FE) estimators for the unbalanced case. Section 3 extends the basic linear model to allow for correlated random slopes. These results allow selection and covariates to be correlated with unobserved heterogeneity that interacts with observed covariates in unbalanced panels.

Section 4 proposes a general method for allowing correlated random effects in nonlinear models. Section 5 discusses the important practical problem of computing partial effects with the heterogeneity averaged out — so called "average partial effects" (APEs). Conveniently, the pooled methods for nonlinear models identify the APEs without restrictions on time series

dependence. We can use the same averaging out of sufficient statistics that is used with balanced panels. Section 6 works through the example of a probit response function, with applications to binary and fractional response. Simple tests for violation of the ignorability of selection are discussed in this section.

Section 7 contains a proposal for comparing fit across different models. The approach appears to be new and provides a unifying framework for choosing among different models with unobserved heterogeneity. Section 8 summarizes some limitations of the current paper and suggests some directions for future research.

2. The linear model with additive heterogeneity

It is useful to begin with the standard linear model with additive heterogeneity. We can set the framework for more complicated settings and at the same time obtain new results that are particularly useful for testing key assumptions. In particular, we obtain a variable addition version of the Hausman (1978) test comparing random effects and fixed effects on the unbalanced panel.

Assume that an underlying population consists of a large number of units for whom data on T time periods are potentially available. We assume random sampling from this population, and let i denote a random draw. Along with the outcome, y_{it}, are covariates, $\mathbf{x}_{it}$. We also draw unobservables for each i; we are particularly interested in unobserved heterogeneity, c_i, which is a scalar for now.

To allow for unbalanced panels, we explicitly introduce a series of selection indicators for each i, $\{s_{i1}, \ldots, s_{iT}\}$, where $s_{it} = 1$ if time period t for unit i can be used in estimation. In this paper, we only use information on units where a full set of data are observed. Therefore, $s_{it} = 1$ if and only if $(\mathbf{x}_{it}, y_{it})$ is fully observed; otherwise, $s_{it} = 0$. In other words, s_{it} indicates whether we have a "complete case" for unit i in time period t. This complete cases scenario is very common in panel data applications with unbalanced panels. In fact, it is the default in software packages that have built in panel data commands.

The linear model with additive heterogeneity is

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \ t = 1, \ldots, T, \tag{2.1}$$

where the $1 \times K$ vector $\mathbf{x}_{it}$ can include a full set of time dummies, or other aggregate time variables. We view this as the equation that holds in an underlying population for all T time periods. We are interested in estimators of $\boldsymbol{\beta}$ that allow for correlation between c_i and the history of covariates, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$. With balanced panels, a common assumption is strict exogeneity of the covariates with respect to the idiosyncratic errors, which leads to the well-known fixed effects estimator and variants. With an unbalanced panel, the key assumption is most easily stated as

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{s}_i) = 0, \ t = 1, \ldots, T, \tag{2.2}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ and $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iT})$ are the histories of the covariates and selection indicators, respectively. Assumption (2.2) implies that observing a data point in any time period cannot be systematically related to the idiosyncratic errors, u_{it}. It is a version of strict exogeneity of selection (along with strict exogeneity of the covariates) conditional on c_i. As a practical matter, (2.2) allows selection s_{it} at time period t to be arbitrarily correlated with $(\mathbf{x}_i, c_i)$, that is, with the observable covariates and the unobserved heterogeneity. For later comparisons with nonlinear models, we can combine (2.1) and (2.2) as

$$E(y_{it}|\mathbf{x}_i, c_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_i, c_i) = \mathbf{x}_{it}\boldsymbol{\beta} + c_i, \tag{2.3}$$

which means we can start from an assumption about a conditional expectation involving the response variable, as is crucial for nonlinear models.

It is well-known – see, for example, Verbeek and Nijman (1996), Hayashi (2001), and Wooldridge (2010, Chapter 17) – that the fixed effects (within) estimator on the unbalanced panel is generally consistent under (2.3), provided there is sufficient time variation in the covariates and the selected sample is not "too small."

One way to characterize the FE estimator on the unbalanced panel is to multiply Eq. (2.1) through by the selection indicator to get

$$s_{it}y_{it} = s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}c_i + s_{it}u_{it}, \ t = 1, \ldots, T. \tag{2.4}$$

Averaging this equation across t for each i gives

$$\bar{y}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \bar{u}_i, \ t = 1, \ldots, T, \tag{2.5}$$

where $\bar{y}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ is the average of the selected observations and $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods observed for unit i. The other averages in (2.5) are defined similarly. If we now multiply (2.5) by s_{it} and subtract from (2.4) we remove c_i:

$$s_{it}(y_{it} - \bar{y}_i) = s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + s_{it}(u_{it} - \bar{u}_i). \tag{2.6}$$

Now we can apply pooled OLS to this equation to obtain the FE estimator on the unbalanced panel. It is straightforward to show that (2.2), along with a rank condition, are sufficient for consistency for fixed T, $N \to \infty$.

As a computational point, note that the time averages of y_{it} and $\mathbf{x}_{it}$ are computed only for time periods where data exist on the *full* set of variables $(\mathbf{x}_{it}, y_{it})$: s_{it} is defined as a complete cases indicator. Consequently, in some applications there are

often pairs (i, t) where we observe some elements in $(\mathbf{x}_{it}, y_{it})$ but where the information on these variables is not used in estimation. The FE estimator is a complete cases estimator.

Because the FE estimator on the unbalanced panel maintains strict exogeneity of the covariates and selection with respect to $\{u_{it} : t = 1, \ldots, T\}$, it is useful to have simple tests of the null hypothesis. When we are especially concerned about feedback, we can add $s_{i,t+1}$ or $(s_{i,t+1}, s_{i,t+1}\mathbf{x}_{i,t+1})$ and use standard fixed effects estimation on the unbalanced panel (dropping time period T). Or, we could add the number of time periods observed after time t, say $r_{it} = s_{i,t+1} + \cdots + s_{iT}$. Another possibility is to add $(s_{i,t-1}, s_{i,t-1}\mathbf{x}_{i,t-1})$ if we want to look for dynamic misspecification of the model. In all cases the add variables should be jointly insignificant.

In the balanced case, it has been known for some time – see Mundlak (1978) – that the FE estimator can be computed as a pooled OLS estimator using the original data and adding the time averages of the covariates as additional explanatory variables. Conveniently, this algebraic result carries over to the unbalanced case. In particular, let $\bar{\mathbf{x}}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}$ be the average of the covariates over the time periods where we observe a complete set of data. The Mundlak device involves estimating the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + v_{it} \tag{2.7}$$

by pooled OLS using the $s_{it} = 1$ observations. Below we provide a general result that implies the coefficient vector $\hat{\boldsymbol{\beta}}$ is identical to the fixed effects (within) estimator on the unbalanced panel. As a cautionary note, in obtaining the FE estimator, any aggregate time variables – in particular, time dummies – should be part of $\mathbf{x}_{it}$, and their time averages must be included in $\bar{\mathbf{x}}_i$. To be more precise, partition $\mathbf{x}_{it}$ as $\mathbf{x}_{it} = (\mathbf{g}_t, \mathbf{h}_{it})$, where $\mathbf{g}_t$ includes the aggregate time effects. In the balanced case, we would not include the time averages $\{\mathbf{g}_t : t = 1, \ldots, T\}$ because that would just be a vector of constants. But in the unbalanced case, the time averages are $T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{g}_t$, and these now vary across i when different time periods are missing for different units. Modern software makes it straightforward to include the time averages of all time-varying variables, including the aggregate time effects.

The point of the previous discussion is that, as in the balanced case, a particularly simple CRE estimator – which models the relationship between c_i and $\{(s_{i1}, s_{i1}\mathbf{x}_{i1}), (s_{i2}, s_{i2}\mathbf{x}_{i2}), \ldots, (s_{iT}, s_{iT}\mathbf{x}_{iT})\}$ as a linear function of $\bar{\mathbf{x}}_i$ – turns out to be very robust: $\boldsymbol{\beta}$ is consistently estimated for *any* dependence between c_i and $\{(s_{i1}, s_{i1}\mathbf{x}_{i1}), (s_{i2}, s_{i2}\mathbf{x}_{i2}), \ldots, (s_{iT}, s_{iT}\mathbf{x}_{iT})\}$.

It is useful to have a general result that contains algebraic equivalences for random effects estimation as well as pooled OLS. Recall that for a model with response variable y_{it} and covariates $(\mathbf{x}_{it}, \mathbf{z}_i)$, where $\mathbf{z}_i$ contains unity and any other set of time-constant variables, and $\mathbf{x}_{it}$ contains any aggregate time variables, the RE estimator can be obtained from the pooled OLS regression

$$y_{it} - \theta_i\bar{y}_i \text{ on } \mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i,\ (1-\theta_i)\bar{\mathbf{x}}_i,\ (1-\theta_i)\mathbf{z}_i \text{ if } s_{it} = 1, \tag{2.8}$$

where $\theta_i = 1 - [\sigma_u^2/(\sigma_u^2 + T_i\sigma_c^2)]^{1/2}$ is a function of T_i and the variance parameters; see, for example, Baltagi (2001, Section 9.2). (In practice, the variance parameters are replaced with estimates, but that is unimportant for an algebraic equivalence.) For our purposes, we do not care where θ_i comes from provided $0 \le \theta_i \le 1$ for all. Pooled OLS ($\theta_i = 0$ for all i), fixed effects ($\theta_i = 1$ for all i), and random effects are special cases. We are certainly not making any assumptions on the underlying model; the following result is purely algebraic.

Proposition 2.1. *Consider the pooled OLS regression in (2.8), where the time averages are computed using the selected observations. Let $\tilde{\boldsymbol{\beta}}$ be the $K \times 1$ vector of coefficients on $\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i$. Then $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FE}$, the fixed effects estimate on the unbalanced panel.* □

The proof is given in the Appendix.

The conclusion of Proposition 2.1 shows that both pooled OLS and RE applied to the Mundlak equation reproduce fixed effects in the general unbalanced case. We can include any time-constant variables in $\mathbf{z}_i$ and we still obtain the FE estimator of $\boldsymbol{\beta}$. This algebraic equivalence allows us to estimate coefficients on time-constant variables while reproducing the FE estimates on the time-varying variables. If desired, we can use the result to detect whether selection is correlated with c_i by including mutually exclusive and exhaustive dummies indicating different values of T_i. Of course, fixed effects are robust to such correlation, but it may be of interest to know whether time-constant indicators of selection help to predict y_{it}. In any case, the coefficients on $\mathbf{x}_{it}$ will not change: we always obtain $\hat{\boldsymbol{\beta}}_{FE}$.

Proposition 2.1 also has useful implications for the balanced panel case concerning Chamberlain's 1982 extension of the Mundlak (1978) device. Without time-constant covariates (only for simplicity), the Chamberlain estimating equation can be written as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha + \mathbf{x}_i\boldsymbol{\lambda} + r_i + u_{it},$$

whereas Mundlak imposes $\boldsymbol{\lambda}_t = \boldsymbol{\xi}/T$, $t = 1, \ldots, T$. But we can rewrite the Chamberlain equation as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha + \bar{\mathbf{x}}_i\boldsymbol{\omega} + \mathbf{x}_{i2}(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1) + \cdots + \mathbf{x}_{iT}(\boldsymbol{\lambda}_T - \boldsymbol{\lambda}_1) + r_i + u_{it},$$

where $\boldsymbol{\omega} = \boldsymbol{\lambda}_1/T$. It now follows from Proposition 2.1, with $s_{it} \equiv 1$ for all (i, t) and $\mathbf{z}_i = (\mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$, that estimating Chamberlain's equation by random effects gives $\hat{\boldsymbol{\beta}}_{FE}$ as the coefficients on $\mathbf{x}_{it}$. This algebraic equivalence (in the balanced case) coheres with the fact that efficiency improvements using the Chamberlain device (over Mundlak) are possible only

when the random effects variance–covariance constant variance and equi-correlation structure do not hold (conditional on the covariates). Abrevaya (2013, Proposition 2.1), obtained the Chamberlain–Mundlak equivalence result for pooled OLS using a different approach.Proposition 2.1 shows that the equivalence holds for the RE estimator, too.

As shown in Abrevaya (2013), extending the Chamberlain approach to the unbalanced case is not so straightforward, and linearly projecting c_i onto $(s_{i1}\mathbf{x}_{i1}, \ldots, s_{iT}\mathbf{x}_{iT})$ results in inconsistent estimation. Yet the Mundlak approach, which nominally seems more restrictive, leads to the fully robust fixed effects estimator.

A particularly useful application of Proposition 2.1 is to provide a simple, regression-based, fully robust Hausman test that effectively compares RE and FE using unbalanced panels. Write a model with time-constant variables $\mathbf{z}_i$ as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + c_i + u_{it}, \ t = 1, \ldots, T, \tag{2.9}$$

where, again, we use a data point if $s_{it} = 1$. Assume that $\mathbf{z}_i$ includes a constant and any other observed time-constant variables that we should include to account for observed heterogeneity. If we use the Mundlak equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \mathbf{z}_i\boldsymbol{\gamma} + a_i + u_{it} \tag{2.10}$$

and estimate this by RE, we know from Proposition 2.1 that the estimate of $\boldsymbol{\beta}$ is the FE estimate. If we impose $\boldsymbol{\xi} = \mathbf{0}$ we obtain the RE estimator on the unbalanced panel. Therefore, the regression-based Hausman test for deciding whether to reject RE in favor of FE is just a fully robust Wald test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$ after RE estimation of (2.10). Any unit with $T_i = 1$ can be included in the estimation and testing, but if we only have units with $T_i = 1$ then $\bar{\mathbf{x}}_i = \mathbf{x}_{it}$ for the single time period t with $s_{it} = 1$, and then $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ cannot be distinguished.

Eq. (2.10) sheds further light on recent work by Guggenberger (2010) on the pre-testing problem in using the Hausman approach to choose between RE and FE estimation. Guggenberger (2010) shows that the size distortions on inference after using the original, nonrobust Hausman test can be severe. Understanding that a fully robust Hausman test is obtained as a test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$ shows that Guggenberger's setup is essentially the same as the problem of pre-testing on sets of regressors. For the Hausman test, the question is whether to include $\bar{\mathbf{x}}_i$: If these are included, the estimate of $\boldsymbol{\beta}$ is FE; if they are not, the estimate is RE. Because the Mundlak approach can be made robust to serial correlation and heteroskedasticity, it is evident that Guggenberger's findings have nothing to do with his studying the nonrobust form of the Hausman test.

A related point is that Eq. (2.10) makes it clear why RE tends to be more efficient than FE, even if the ideal assumptions for RE — serial independence and homoskedasticity of $\{u_{it}\}$ — fail. FE estimation of $\boldsymbol{\beta}$ is the same as including $\bar{\mathbf{x}}_i$ in (2.10) while RE estimation is dropping $\bar{\mathbf{x}}_i$. By construction, $\bar{\mathbf{x}}_i$ is collinear with $\mathbf{x}_{it}$, and the degree of collinearity increases as the time variation in $\{\mathbf{x}_{it}\}$ decreases. So FE includes collinear variables while RE does not. It is well known from standard regression theory that dropping collinear variables when they have zero coefficients – that is, $\boldsymbol{\xi} = \mathbf{0}$ – can produce substantial efficiency gains. That there might be serial correlation or heteroskedasticity in $\{u_{it}\}$, or an unbalanced panel, does not materially influence the argument.

We can also use the Mundlak CRE formulation to test a subset of coefficients in $\boldsymbol{\xi}$. For example, perhaps we are interested in x_{ti1} (which may be a policy variable) and it does not have much time variation, leading to a large standard error for the FE estimate. To test whether $\{x_{it1}\}$ is exogenous with respect to c_i, after controlling for $(\bar{x}_{i2}, \ldots, \bar{x}_{iK}, \mathbf{z}_i)$, we can test $H_0 : \xi_1 = 0$, where ξ_1 is the coefficient on $\bar{x}_{i1}$, using a fully robust t statistic. A failure to reject provides some justification for estimating the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \xi_2\bar{x}_{i2} + \cdots + \xi_K\bar{x}_{iK} + \mathbf{z}_i\boldsymbol{\gamma} + a_i + u_{it} \tag{2.11}$$

by random effects — subject to the usual pre-testing problem. Dropping $\bar{x}_{i1}$ could lead to a substantially more precise estimate of β_1 (the coefficient on x_{it1}). One might want to include dummies for the different values of T_i to better control for selection bias.

In summary, this section has shown that even if we nominally make the strong assumption

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, E\,(a_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}) = 0,$$

the resulting CRE estimator – whether implemented by pooled OLS or RE – is identical to fixed effects, which puts no restrictions on $E(c_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\})$. This extension of the usual (Mundlak, 1978) result to the unbalanced case has several useful applications, as described above.

3. Linear models with correlated random slopes

Now we consider a linear model that has unit-specific slopes. Wooldridge (2005) shows that using the usual FE estimator in a linear model where the random slopes are ignored has some robustness properties for estimating the average partial effects. But those findings do not carry over to unbalanced panels where selection may be correlated with heterogeneity: the slope heterogeneity becomes part of the error term, and correlation between selection and the heterogeneity generally causes inconsistency.

To derive methods for unbalanced panels, state the model as

$$E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = a_i + \mathbf{x}_{it}\mathbf{b}_i, \tag{3.1}$$

so, in the population, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is strictly exogenous conditional on $(a_i, \mathbf{b}_i)$. Define $a_i = \alpha + c_i$, $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$ and write

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \mathbf{x}_{it}\mathbf{d}_i + u_{it}, \tag{3.2}$$

where $E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = E(u_{it}|\mathbf{x}_i, c_i, \mathbf{d}_i)$ for all t. We allow that selection may be related to $(\mathbf{x}_i, a_i, \mathbf{b}_i)$ but not the idiosyncratic shocks:

$$E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) \tag{3.3}$$

or

$$E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = 0, \ \ t = 1, \ldots, T, \tag{3.4}$$

which is an obvious extension of assumption (2.3).

In what follows, we allow all elements of $\mathbf{b}_i$ to be heterogeneous. If we had only a few such elements, and a sufficient number of time periods, we could proceed by eliminating those elements of $\mathbf{b}_i$ via a generalized within transformation and then proceed with estimation of the constant slopes. Such an approach would be the unbalanced version of the methods described by Wooldridge (2010, Chapter 11). This approach is attractive in specific instances, but it cannot be used in general. Setting some elements of $\mathbf{b}_i$ to constants in applying the subsequent approach is straightforward.

To study estimation on an unbalanced panel, multiply (3.2) through by the selection indicator:

$$s_{it}y_{it} = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}c_i + s_{it}\mathbf{x}_{it}\mathbf{d}_i + s_{it}u_{it} \tag{3.5}$$

We handle the presence of intercept and slope heterogeneity by conditioning on the entire history of selection and the values of the covariates if selected, $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. If $s_{it} = 0$ the observation is not used; if $s_{it} = 1$ the observation is used and we observe $\mathbf{x}_{it}$. It might seem better to condition on $\{(\mathbf{x}_{i1}, s_{i1}), (\mathbf{x}_{i2}, s_{i2}), \ldots, (\mathbf{x}_{iT}, s_{iT})\}$, but if, say, the heterogeneity depends only on the history of covariates, we would be left with an equation that is not estimable unless the covariates are always observed. Therefore, to obtain a true estimating equation, we condition on $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$.

For notational simplicity, write $\mathbf{h}_i \equiv \{\mathbf{h}_{it} : t = 1, \ldots, T\} \equiv \{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. Then, extending Mundlak (1978) and Chamberlain (1982, 1984), we work with

$$E(s_{it}y_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}E(c_i|\mathbf{h}_i) + s_{it}\mathbf{x}_{it}E(\mathbf{d}_i|\mathbf{h}_i) \tag{3.6}$$

and then make assumptions concerning $E(c_i|\mathbf{h}_i)$ and $E(\mathbf{d}_i|\mathbf{h}_i)$. Actually, because we can eliminate c_i using the within transformation, we could just focus on $E(\mathbf{d}_i|\mathbf{h}_i)$. However, assuming we know models for $E(\mathbf{d}_i|\mathbf{h}_i)$ but not for $E(c_i|\mathbf{h}_i)$ is somewhat arbitrary, and so we first consider the case where we model all expectations.

It is useful to point out that if $(a_i, \mathbf{b}_i)$ are assumed to be independent (or, at least, mean independent) of $\{(\mathbf{x}_{i1}, s_{i1}), (\mathbf{x}_{i2}, s_{i2}), \ldots, (\mathbf{x}_{iT}, s_{iT})\}$ – an assumption often implicit in random coefficient frameworks – then the issue of how to model $E(c_i|\mathbf{h}_i)$ and $E(\mathbf{d}_i|\mathbf{h}_i)$ disappears. The term $s_{it}E(c_i|\mathbf{h}_i) + s_{it}\mathbf{x}_{it}E(\mathbf{d}_i|\mathbf{h}_i)$ would be identically zero, which means we would be left with $E(s_iy_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta}$. Pooled estimation using the selected sample or generalized least squares methods can be applied.

A simple approach to allowing $E(a_i, \mathbf{b}_i|\mathbf{h}_i)$ to depend on $\mathbf{h}_i$ is to model the expectations as functions of $\{\mathbf{h}_{it} : t = 1, \ldots, T\}$. In the balanced panel case, Altonji and Matzkin (2005) suggested using exchangeable functions in a fully nonparametric setting. The leading examples of exchangeable functions are averages. In keeping with the motivation from Section 2, we might choose

$$\mathbf{w}_i \equiv (T_i, \bar{\mathbf{x}}_i) \tag{3.7}$$

as exchangeable functions satisfying

$$E(c_i|\mathbf{h}_i) = E(c_i|\mathbf{w}_i), E(\mathbf{d}_i|\mathbf{h}_i) = E(\mathbf{d}_i|\mathbf{w}_i). \tag{3.8}$$

But we could use nonexchangeable forms, too, such as the Chamberlain device: $\mathbf{w}_i = (s_{i1}, \ldots, s_{iT}, s_{i1}\mathbf{x}_{i1}, \ldots, s_{iT}\mathbf{x}_{iT})$. Or, we could use unit-specific trends, or variances, and so on.

Unlike in Section 2, we now have to take our assumptions on $E(c_i|\mathbf{h}_i)$ and $E(\mathbf{d}_i|\mathbf{h}_i)$ seriously. In the correlated random slopes model with unbalanced panels, there are no known robustness results if the conditional mean restrictions fail to hold. The hope is that, because the Mundlak device for the model in Section 2 is fully robust, using similar assumptions for heterogeneous slope models will tend to work well. But it must be emphasized that this is an argument based on intuition rather than rigor. Nevertheless, with small T and many heterogeneous slopes, it is unclear how else one would proceed while allowing the slopes to be correlated with the covariates and selection.

Focusing for now on (3.7), if we assume that these expectations depend only on $\bar{\mathbf{x}}_i$, and in a linear fashion, then

$$E(c_i|\mathbf{h}_i) = (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i})\boldsymbol{\xi} \tag{3.9}$$

$$E(\mathbf{d}_i|\mathbf{h}_i) = [(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i}) \otimes \mathbf{I}_K]\boldsymbol{\eta}, \tag{3.10}$$

where K is the dimension of $\mathbf{x}_{it}$ and $\boldsymbol{\mu}_{\bar{\mathbf{x}}_i} = E(\bar{\mathbf{x}}_i)$ is subtracted from $\bar{\mathbf{x}}_i$ to impose zero unconditional means on c_i and $\mathbf{d}_i$. Note that $(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i}) \otimes \mathbf{I}_K$ is a $K \times K^2$ matrix and $\boldsymbol{\eta}$ is a $K^2 \times 1$ vector. If we insert these expectations into (3.6) and use simple algebra, we obtain

$$E(s_{it}y_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i})\boldsymbol{\xi} + s_{it}[(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i}) \otimes \mathbf{x}_{it}]\boldsymbol{\eta}, \tag{3.11}$$

which is an equation with the time averages and each time average interacted with each time-varying covariate. It is now obvious that we can use pooled OLS on the selected sample to consistently estimate α, $\boldsymbol{\beta}$ (the main vector of interest), $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$. We can even use, say, random effects estimation, but inference should be made robust to arbitrary heteroskedasticity and serial correlation. As a practical matter, we replace $\boldsymbol{\mu}_{\bar{\mathbf{x}}_i}$ with $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{x}}_i} = N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i$ as a consistent estimator of $\boldsymbol{\mu}_{\bar{\mathbf{x}}_i}$. Notice that $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{x}}_i}$ is consistent for the quantity we need, which is the expected value of $\bar{\mathbf{x}}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}$.

If we drop the set of interactions $(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i})\otimes\mathbf{x}_{it}$, we know from Section 2 the resulting estimator would be the FE estimator on the unbalanced panel. This suggests a simple test for whether we need to further consider correlation of selection and the random slopes. Estimate the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + [(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}_i})\otimes\mathbf{x}_{it}]\boldsymbol{\eta} + a_i + u_{it} \tag{3.12}$$

by fixed effects, so that a_i is removed without imposing any assumptions on its conditional distribution. If we cannot reject $H_0 : \boldsymbol{\eta} = \mathbf{0}$, we might ignore the possibility of random slopes and just use standard FE estimation on the unbalanced panel.

If we conclude that we need to account for the random slopes, the assumptions in (3.9) and (3.10) might be too restrictive. For one, they assume that T_i does not directly appear in $E(c_i, \mathbf{d}_i|\mathbf{h}_i)$. Second, since $\bar{\mathbf{x}}_i$ is an average using T_i elements, it is possible the coefficients change with T_i. [This certainly would be the case under joint normality given any sequence of selection indicators with sum T_i.] We can allow an unrestricted set of slopes by extending the earlier assumption to

$$E(c_i|\mathbf{h}_i) = E(c_i|T_i, \bar{\mathbf{x}}_i) = \sum_{r=1}^{T}\psi_r\{1[T_i = r] - \rho_r\} + \sum_{r=1}^{T} 1[T_i = r]\cdot(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r)\boldsymbol{\xi}_r \tag{3.13}$$

$$E(\mathbf{d}_i|\mathbf{h}_i) = E(\mathbf{d}_i|T_i, \bar{\mathbf{x}}_i) = \sum_{r=1}^{T}\{1[T_i = r] - \rho_r\}\boldsymbol{\kappa}_r + \sum_{r=1}^{T} 1[T_i = r]\cdot[(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r)\otimes\mathbf{I}_K]\boldsymbol{\eta}_r, \tag{3.14}$$

where the $\boldsymbol{\mu}_r$ are the expected values of $\bar{\mathbf{x}}_i$ given r time periods observed and ρ_r is the fraction of observations with r time periods:

$$\boldsymbol{\mu}_r = E(\bar{\mathbf{x}}_i|T_i = r),\ \ \rho_r = E\{1[T_i = r]\} \tag{3.15}$$

As a practical matter, the formulation in (3.13) and (3.14) is identical to running separate regressions for each T_i:

$$y_{it} \text{ on } 1,\ \mathbf{x}_{it},\ \bar{\mathbf{x}}_i\,, (\bar{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_r)\otimes\mathbf{x}_{it},\ \text{ for } s_{it} = 1 \tag{3.16}$$

where $\hat{\boldsymbol{\mu}}_r = N_r^{-1}\left(\sum_{i=1}^{N} 1[T_i = r]\bar{\mathbf{x}}_i\right)$ and N_r is the number of observations with $T_i = r$. The coefficient on $\mathbf{x}_{it}$, $\hat{\boldsymbol{\beta}}_r$, is the APE given $T_i = r$. We can average these across r to obtain the overall APE. There is, however, a cost in allowing the flexibility in (3.13) and (3.14): we cannot identify an APE for $T_i = 1$ unless we set the coefficients on $\bar{\mathbf{x}}_i$ and $(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r)\otimes\mathbf{x}_{it}$ equal to zero. So, we could just exclude the $T_i = 1$ observations from the APE calculations, or we can impose restrictions that we did previously. [The same issue arises if we use fixed effects estimation to obtain a different $\hat{\boldsymbol{\beta}}_r$ for each T_i: we must exclude the $T_i = 1$ subsample.] Under the assumption that $\{(s_{it}, \mathbf{x}_{it}) : t = 1, \ldots, T\}$ is independent and identically distributed, the coefficients in (3.13) and (3.14) are linear functions of T_i, and such a restriction means we can use the $T_i = 1$ observations.

A special case of the previous model is the so-called random trend model, where $\mathbf{x}_{it}$ includes (in the simplest case) t, so that each unit has its own linear trend. Then, we might want to allow the random trend to be correlated with features of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ other than the average and number of observed time periods. For example, for each i we could "estimate" unit-specific intercept and trend coefficient by running regressions

$$s_{it}\mathbf{x}_{it} \text{ on } s_{it},\ s_{it}t, t = 1, \ldots, T, \tag{3.17}$$

and then allow these coefficients to be correlated with c_i and $\mathbf{d}_i$.

As in the case with additive heterogeneity, we have available simple tests of dynamic selection bias. Under the ignorability assumption (3.3), no other functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ should appear in $E(s_{it}y_{it}|\mathbf{h}_i)$. In constructing a test we cannot include s_{it} as an explanatory variable at time t because we only use data with $s_{it} = 1$. We might add $(s_{i,t+1}, s_{i,t+1}\mathbf{x}_{i,t+1})$ or $(s_{i,t-1}, s_{i,t-1}\mathbf{x}_{i,t-1})$ to an estimating equation such as (3.12) and compute a fully robust (to serial correlation and heteroskedasticity) exclusion test.

4. A modeling approach for nonlinear models

We can apply the approach for linear models with random slopes to general nonlinear models. We assume that interest lies in some feature of the distribution

$$D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \tag{4.1}$$

where, in general, $\mathbf{y}_{it}$ can be a vector, $\mathbf{x}_{it}$ is a set of observed conditioning variables, and $\mathbf{c}_i$ is a vector heterogeneity. The strict exogeneity assumption is

$$D(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{c}_i) = D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \tag{4.2}$$

where $\mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}\}$ is the entire history of covariates. We assume that we have specified, for each t, a density for $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$. This density is written with placeholder arguments as $g_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a set of finite dimensional parameters. Here we focus on the case of specifying marginal distributions for each t, rather than a joint distribution. Pooled methods are generally more robust because they do not restrict dependence over time. Plus, as discussed in Wooldridge (2010), average partial effects are generally identified by pooled estimation methods, and computationally they are relatively simple.

Given the strict exogeneity assumption, selection is assumed to be ignorable conditional on $(\mathbf{x}_i, \mathbf{c}_i)$:

$$D(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i) = D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i), t = 1, \ldots, T. \tag{4.3}$$

As in the case of linear models, (4.3) allows selection to be arbitrarily correlated with $(\mathbf{x}_i, \mathbf{c}_i)$ but not generally with "shocks" to $\mathbf{y}_{it}$.

Our correlated random effects approach specifies a model for

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}). \tag{4.4}$$

Let $\mathbf{w}_i$ be a vector of known functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ that act as sufficient statistics, so that

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}) = D(\mathbf{c}_i|\mathbf{w}_i), \tag{4.5}$$

which is the distributional version of the conditional mean assumption in Section 3.

Now, because $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i, s_{it} = 1) = D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, it follows that the density of $\mathbf{y}_{it}$ given $(s_{it}, s_{it}\mathbf{x}_{it}, \mathbf{c}_i)$ is $g_t(\mathbf{y}_t|s_{it}\mathbf{x}_{it}, \mathbf{c}_i; \boldsymbol{\gamma})$ when $s_{it} = 1$. As we are only using data with $s_{it} = 1$, this is enough to construct the density used in estimation: the one conditional on $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. Let $h(\mathbf{c}|\mathbf{w}_i; \boldsymbol{\delta})$ be a parametric density for $D(\mathbf{c}_i|\mathbf{w}_i)$. Then the density we need (again for $s_{it} = 1$) is

$$f_t(\mathbf{y}_t|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}) = \int_{\mathbb{R}^M} g_t(\mathbf{y}_t|\mathbf{x}_{it}, \mathbf{c}; \boldsymbol{\gamma}) h(\mathbf{c}|\mathbf{w}_i; \boldsymbol{\delta}) d\mathbf{c}, \tag{4.6}$$

where M is the dimension of $\mathbf{c}_i$, and this is obtainable given models $g_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$ and $h(\mathbf{c}|\mathbf{w}; \boldsymbol{\delta})$. In effect, the same calculations used to "integrate out" unobserved heterogeneity in the balanced case can be used here, too.

For each i, a partial log-likelihood function is

$$\sum_{t=1}^{T} s_{it} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})], \tag{4.7}$$

(where we abuse notation by not distinguishing the true parameters from a generic value). The true values of the parameters maximize $E[\log f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})]$ given $s_{it} = 1$, and so the partial MLE generally identifies $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$. The partial log likelihood for the full sample is

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})]. \tag{4.8}$$

The large-N, fixed-T asymptotics falls into the general M-estimation framework described in Wooldridge (2010, Chapter 12).

With a pooled method, inference needs to be made robust to the serial dependence in the scores from Eq. (4.8). Let $\boldsymbol{\theta}$ be the vector of all parameters, and assume identification holds along with regularity conditions. Define the scores and Hessians as

$$\mathbf{r}_{it}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\theta})]'$$
$$\mathbf{H}_{it}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\theta})]$$

Then

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

where

$$\mathbf{A} = -E\left[\sum_{t=1}^{T} s_{it}\mathbf{H}_{it}(\boldsymbol{\theta})\right]$$

$$\mathbf{B} = Var\left[\sum_{t=1}^{T} s_{it}\mathbf{r}_{it}(\boldsymbol{\theta})\right] = E\left\{\left[\sum_{t=1}^{T} s_{it}\mathbf{r}_{it}(\boldsymbol{\theta})\right]\left[\sum_{t=1}^{T} s_{it}\mathbf{r}_{it}(\boldsymbol{\theta})\right]'\right\}$$

The definition of $\mathbf{B}$ allows correlation across the scores for different time periods. Estimators of these matrices are standard: we can replace the expectation with an average across i and replace $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$. Canned software packages that allow for "clustering" in panel data contexts will produce valid standard errors and test statistics.

For many applications it is useful to know that essentially the same arguments carry through when we specify models for conditional means only. That is, if we start with $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ as the object of interest, then we can obtain $E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$ by integrating $m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ with respect to the density of $\mathbf{c}_i$ given $\mathbf{w}_i$. The resulting conditional mean will generally depend on the models $m_t(\mathbf{x}_t, \mathbf{c})$ and $h(\mathbf{c}|\mathbf{w})$ being correctly specified, but if we assume correct specification, we can use a variety of pooled quasi-MLEs for estimation. For example, if y_{it} is a fractional response, we can use the Bernoulli quasi-log likelihood (QLL); if y_{it} is nonnegative, such as a count variable, we can use the Poisson QLL.

5. Estimating average partial effects

In most nonlinear models, the parameters $\boldsymbol{\gamma}$ appearing in $g_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$ provide only part of the story for obtaining the effect of $\mathbf{x}_t$ on $\mathbf{y}_t$. The presence of heterogeneity usually means that the elements of $\boldsymbol{\gamma}$ can, at best, provide directions and relative magnitudes of effects. Fortunately, for the setup developed in Section 4 we have enough information to identify and estimate partial effects with the heterogeneity averaged out.

We follow Blundell and Powell (2003) and define the *average structural function* (ASF) for a scalar response, y_t. Let $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ be the mean function. Then

$$ASF(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] \tag{5.1}$$

is the conditional mean function (as a function of the dummy argument, $\mathbf{x}_t$) with the heterogeneity, $\mathbf{c}_i$, averaged out. Given the ASF, we can compute partial derivatives, or discrete changes, with respect to the elements of $\mathbf{x}_t$. As discussed in Wooldridge (2010, Section 2.2.5), this generally produces the *average partial effects* (APEs), that is, the partial derivatives (or changes) with the heterogeneity averaged out. Fortunately, the ASF (and, therefore, APEs) is often easy to obtain.

Let $q_t(\mathbf{x}_t, \mathbf{w}; \boldsymbol{\theta})$ denote the mean associated with $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{w}; \boldsymbol{\theta})$. Then, as discussed in Wooldridge (2010) for the balanced case,

$$ASF(\mathbf{x}_t) = E_{\mathbf{w}_i}[q_t(\mathbf{x}_t, \mathbf{w}_i; \boldsymbol{\theta})]; \tag{5.2}$$

that is, we can obtain the ASF by averaging out the observed vector of sufficient statistics, $\mathbf{w}_i$, from $E(y_{it}|\mathbf{x}_t, \mathbf{w}_i, s_{it} = 1)$ rather than averaging out $\mathbf{c}_i$ from $E(y_{it}|\mathbf{x}_t, \mathbf{c}_i)$. In leading cases, we have direct estimates of $q_t(\mathbf{x}_t, \mathbf{w}; \boldsymbol{\theta})$, in which case we have a simple, consistent estimator of $ASF(\mathbf{x}_t)$:

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N} q_t(\mathbf{x}_t, \mathbf{w}_i; \hat{\boldsymbol{\theta}}) \tag{5.3}$$

We can use this expression to obtain APEs by taking derivatives or changes with respect to elements of $\mathbf{x}_t$, for example,

$$\widehat{APE}_{tj}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N} \frac{\partial q_t(\mathbf{x}_t, \mathbf{w}_i; \hat{\boldsymbol{\theta}})}{\partial x_{tj}} \tag{5.4}$$

Standard errors of such quantities can be difficult to obtain by the delta method, but the panel bootstrap – where resampling is done in the cross section dimension – is straightforward. further, because we are using pooled methods, the bootstrap is usually quite tractable computationally.

With an unbalanced panel, there is a somewhat subtle point about computing a single average partial effect. Generally, standard software for estimating APEs would compute

$$N^{-1}\sum_{i=1}^{N}\left[T_i^{-1}\sum_{t=1}^{T} s_{it}\frac{\partial q_t(\mathbf{x}_{it}, \mathbf{w}_i; \hat{\boldsymbol{\theta}})}{\partial x_{tj}}\right], \tag{5.5}$$

or the discrete version. But if selection s_{it} depends on $\mathbf{x}_{it}$, averaging across the selected sample does not consistently estimate the average partial effect. We might have to be satisfied with computing the APE in the selected sample. If the data were always available on $\mathbf{x}_{it}$ then we would set $s_{it} = 1$ and $T_i = T$ in (5.5).

6. Example: A probit response function

We now work through a probit response function to show how the proposed method applies to standard models. If y_{it} is a binary response then the model with a single source of heterogeneity and strictly exogenous covariates is

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), t = 1, \ldots, T \tag{6.1}$$

where $\mathbf{x}_{it}$ can include time dummies or other aggregate time variables. We do not restrict the serial dependence in the responses. Once we specify (6.1) and assume that selection is conditionally ignorable for all t, that is,

$$P(y_{it} = 1|\mathbf{x}_i, c_i, \mathbf{s}_i) = P(y_{it} = 1|\mathbf{x}_i, c_i), \tag{6.2}$$

all that is left is to specify a model for $D(c_i|\mathbf{w}_i)$ for suitably chosen functions $\mathbf{w}_i$ of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. As in the linear case, it makes sense to at least initially choose exchangeable functions that extend the usual choices in the balanced case.

For example, we can allow $E(c_i|\mathbf{w}_i)$ to be a linear function of the time averages with different coefficients for each number of periods:

$$E(c_i|\mathbf{w}_i) = \sum_{r=1}^{T} \psi_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r \tag{6.3}$$

Thus, we either have to be content with estimating the APE over the subpopulation with $T_i \geq 2$ or impose more restrictions, such as linear functions in T_i in (6.3). At a minimum we should allow the variance of c_i to change with T_i; a simple yet flexible specification is

$$Var(c_i|\mathbf{w}_i) = \exp\left(\tau + \sum_{r=1}^{T-1} 1[T_i = r]\boldsymbol{\omega}_r\right) \tag{6.4}$$

where τ is the variance for the base group, $T_i = T$, and the each ω_r is the deviation from the base group. If we also maintain that $D(c_i|\mathbf{w}_i)$ is normal, then we obtain the following response probability for $s_{it} = 1$:

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1) = \Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=2}^{T} \psi_r 1[T_i = r] + \sum_{r=2}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r}{\left\{1 + \exp\left(\tau + \sum_{r=1}^{T-1} 1[T_i = r]\omega_r\right)\right\}^{1/2}}\right] \tag{6.5}$$

In the case of the usual model with balanced data, the ω_r are all zero, and then only the coefficients scaled by $[1 + \exp(\tau)]^{1/2}$ are identified. Fortunately, it is exactly these scaled coefficients that determine the average partial effects. A convenient reparameterization is

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{w}_i) = \Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=1}^{T} \psi_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r}{\exp\left(\sum_{r=2}^{T} 1[T_i = r]\omega_r\right)^{1/2}}\right] \tag{6.6}$$

so that the denominator is unity when all ω_r are zero. As an additional bonus, the formulation in (6.6) is directly estimable by so-called "heteroskedastic probit" software, where the explanatory variables at time t are $(1, \mathbf{x}_{it}, 1[T_i = 2] \cdot \bar{\mathbf{x}}_i, \ldots, 1[T_i = T] \cdot \bar{\mathbf{x}}_i)$ and the explanatory variables in the variance are simply the dummy variables $(1[T_i = 2], \ldots, 1[T_i = T - 1])$.

With the estimating equation specified as in (6.6), the average structural function is fairly straightforward to estimate:

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} \Phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T} \hat{\psi}_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_r}{\exp\left(\sum_{r=2}^{T} 1[T_i = r]\hat{\omega}_r\right)^{1/2}}\right] \tag{6.7}$$

where the coefficients with "^" are from the pooled heteroskedastic probit estimation. Notice how the functions of $(T_i, \bar{\mathbf{x}}_i)$ are averaged out, leaving the result a function of $\mathbf{x}_t$. If, say, x_{tj} is continuous, its APE is estimated as

$$\hat{\beta}_j \left\{N^{-1} \sum_{i=1}^{N} \phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T} \hat{\psi}_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_r}{\exp\left(\sum_{r=2}^{T} 1[T_i = r]\hat{\omega}_r\right)^{1/2}}\right]\right\} \tag{6.8}$$

where $\phi[\cdot]$ is the standard normal pdf. This is still a function of $\mathbf{x}_t$. Notice that in the continuous or discrete case, $\hat{\beta}_j$ provides the direction of the effect, but the magnitude of the effect is considerably more complicated (and generally a function of $\mathbf{x}_t$, of course). The parameters of the model for $D(c_i|\mathbf{w}_i)$ appear directly in the ASF and APEs, and so they cannot be considered "nuisance" or "incidental" parameters.

The above procedure applies, without change, if y_{it} is a fractional response; that is, $0 \leq y_{it} \leq 1$. Then, we interpret the original model as $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$, and then partial effects are on the mean response. As is well known – for example, Gourieroux et al. (1984) – the Bernoulli log likelihood is in the linear exponential family, and so it identifies the parameters of a correctly specified conditional mean. Under the assumptions given, we have the correct functional form for $E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$.

We can easily add the interactions $1[T_i = r] \cdot \bar{\mathbf{x}}_i$ to the variance function for added flexibility; if we maintain conditional normality of the heterogeneity, we are still left with an estimating equation of the heteroskedastic probit form. As in (6.7) and (6.8), those extra functions of $(T_i, \bar{\mathbf{x}}_i)$ get averaged out in computing APEs.

The normality assumption, as well as specific functional forms for the mean and variance, might seem restrictive. An important practical point is that, once we know the APEs are identified by averaging $\mathbf{w}_i$ out of $q_t(\mathbf{x}_t, \mathbf{w}_i, \boldsymbol{\theta}) = E[m_t(\mathbf{x}_t, \mathbf{c}_i)|\mathbf{w}_i]$, we are free to use any number of approximations to the true distribution. For example, one could use a logit functional form rather than probit — even though that particular response probability cannot be easily derived from an underlying model for $E(y_{it}|\mathbf{x}_{it}, c_i)$.

Perhaps more useful is extending the functional form inside the probit function. Because we probably should allow different coefficients for each T_i, the notation gets complicated, but we can add interactions of the form

$$1[T_i = r] \cdot (\bar{\mathbf{x}}_i \otimes \mathbf{x}_{it}). \tag{6.9}$$

This is in the spirit of allowing random slopes on $\mathbf{x}_{it}$ in the original probit specification, but this particular estimating equation would not be easily derivable from such a model. Instead, as in Blundell and Powell (2003), it recognizes that quantities of interest can be obtained without even specifying a particular model for $E(y_{it}|\mathbf{x}_{it}, c_i)$.

As discussed in Sections 3 and 4 for linear models, we can easily relax the restriction that $D(c_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\})$ depends only on $(T_i, \bar{\mathbf{x}}_i)$. We can use a Chamberlain device, sample variances and covariances, individual-specific trends, or break the time period into intervals and use averages over those intervals.

Under the key ignorability of selection assumption (4.3), one can justify estimation on any balanced subset of data, including the panel with only a complete set of time periods ($T_i = T$). At a minimum, restricting attention to the largest balanced subpanel – using standard CRE methods for balanced panels – can be used as a robustness check.

As in the linear case, we can easily test for dynamic forms of selection bias by including, say, $s_{i,t+1}$ and $s_{i,t+1}\mathbf{x}_{i,t+1}$ in any of the estimations and obtain fully robust joint Wald tests.

Everything just covered for the probit (fractional probit) case extends to ordered probit. Further, one can use some new, computationally simple strategies for handling multinomial responses. Let y_{it} be an unordered multinomial response. Then, rather than specifying $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ to have any specify form, we can move directly to specifications for $D(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$ for $\mathbf{w}_i$ the chosen sufficient statistics of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. For example, we might just estimate multinomial logit for $D(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$, or nested logit, or some other relatively simple model. Then, the APEs for the response probabilities are obtained by averaging out $\mathbf{w}_i$. In fact, the multinomial quasi-MLE can be applied when the y_{it} are shares summing to one, again relying on Gourieroux, Monfort, and Trognon (1984).

7. Proposals for goodness of fit

An issue that arises in comparing different models with unobserved heterogeneity is how one measures goodness of fit. Measuring fit is further complicated when different estimation methods are used. For example, suppose that y_{it} is binary or a fractional response, and we want to compare a linear model – with just a single, additive heterogeneity – to a binary or fractional response model, also with a single source of heterogeneity. If the linear model is estimated by fixed effects and the probit/fractional model uses the methods proposed in Section 6, it is not clear how one can determine which model fits best, and whether the functional form and distributional assumptions imposed in the fractional case are contributing to a poor fit.

The correlated random effects setting allows us to compare different models using readily available goodness-of-fit measures. Suppose initially that there is no missing data problem and let $\mathbf{w}_i$ be the functions of $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}\}$ such that $D(\mathbf{c}_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i|\mathbf{w}_i)$ is assumed. The partial MLE approach implies densities for the conditional distributions $D(y_{it}|\mathbf{x}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$, which we denoted $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{w}; \boldsymbol{\gamma}, \boldsymbol{\delta})$. Thus, to compare fit across models where densities $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{w}; \boldsymbol{\gamma}, \boldsymbol{\delta})$ are implied, we can use the value of the partial log likelihood evaluated at the partial MLEs $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})]. \tag{7.1}$$

Using (7.1) as a measure of fit combines the "structural" density $g_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$ and the density modeling $D(\mathbf{c}_i|\mathbf{x}_i)$, $h(\mathbf{c}|\mathbf{w}; \boldsymbol{\delta})$ — including which functions of $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\}$ are allowed to be related to $\mathbf{c}_i$. For a given $g_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$, different choices for $h(\mathbf{c}|\mathbf{w}; \boldsymbol{\delta})$ are easily compared.

Using the partial log likelihood to measure fit extends directly to the unbalanced case. The partial log likelihood is evaluated for the observed sample, which means inserting an s_{it} into (7.1). And $\mathbf{w}_i$ is now a function of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$, such as $(T_i, \bar{\mathbf{x}}_i)$.

As discussed in Wooldridge (2010, Section 13.11.2), Vuong's 1989 model selection test can be extended to partial MLEs, provided one is careful in computing a standard error. Applying Vuong's approach to unbalanced panels under the ignorability-of-selection condition (4.3) poses no difficulties.

In many cases we are mainly interested in the fit of the conditional mean. As in Section 5, let $q_t(\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})$ be $E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$. Then we can compute a sum of squared residuals as

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}[y_{it} - q_t(\mathbf{x}_{it}, \mathbf{w}_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})]^2. \tag{7.2}$$

This measure of fit is comparable across models that differ in $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ or $D(\mathbf{c}_i|\mathbf{w}_i)$, including the choice of $\mathbf{w}_i$. For example, if y_{it} is, say, a binary or fractional response, we can compare a linear CRE model to a CRE binary or fractional response model by using the mean functions

$$\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \sum_{r=2}^{T-1} 1[T_i = r]\omega_r$$

for the linear case and, say,

$$\Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta}+\sum_{r=2}^{T}\psi_r 1[T_i=r]+\sum_{r=2}^{T}1[T_i=r]\cdot\bar{\mathbf{x}}_i\boldsymbol{\xi}_r}{\exp\left(\sum_{r=1}^{T-1}1[T_i=r]\omega_r\right)^{1/2}}\right]$$

for the nonlinear case. The estimate of $\boldsymbol{\beta}$ in the linear model is the FE estimator, as shown in Proposition 2.1. As in Vuong (1989), we can add penalties to the sum of squared residuals for number of parameters if desired.

8. Concluding remarks

I have offered some simple strategies for allowing unbalanced panels in correlated random effects models. The key requirement of the approach is to model $D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\})$. I have focused on Mundlak-type assumptions, but more flexible Chamberlain-type projections can be used, too. Also, while I have imposed parametric assumptions (although serial independence is entirely unrestricted), the approaches of Blundell and Powell (2003) and Altonji and Matzkin (2005) can be applied to obtain estimators less sensitive to parametric assumptions.

A general charge leveled at parametric CRE approaches is that having to model $D(\mathbf{c}_i|\mathbf{w}_i)$ means that we may face logical inconsistencies when we think of adding another time period to the data set. For example, even in the balanced case, the restrictions on the covariate process such that $D(\mathbf{c}_i|\mathbf{w}_i)$ is normal for any T are quite strong. This is a valid criticism of the CRE approach, and it motivates the research on fixed effects approaches to nonlinear unobserved effects models discussed in Section 1. It also gives impetus to pursuing the nonparametric and semiparametric CRE approaches.

Having unbalanced panels creates even more logical inconsistencies if we try linking conditional distributions across different numbers of time periods. Nevertheless, empirical researchers ignore essentially the same logical inconsistencies on a daily basis. Whenever, say, a new covariate is added to a probit model, the new model cannot be a probit model if the original model was, unless the new covariate is essentially normally distributed.

In some settings, it is difficult to know what one would do other than a CRE approach. For example, if we start with a probit model with lots of heterogeneity, say $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$, the CMLE approach does not apply, and to treat $(a_i, \mathbf{b}_i)$ as parameters to estimate requires $T_i \geq K + 1$. In practice, T_i should be much larger than $K + 1$, or the incidental parameters problem will be severe. By contrast, a CRE approach is tractable, and has known large-N, fixed-T properties if we properly model $D(\mathbf{c}_i|\mathbf{w}_i)$. Moreover, if we simply start with very flexible models for $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$ – as discussed in Section 6 – and then average out $\mathbf{w}_i$, we can approximate the APEs. How well we do depends on many factors. Sophisticated simulation studies can certainly help us understand the tradeoffs.

The assumption of strictly exogenous covariates is strong and needs to be relaxed. Relaxing strict exogeneity poses challenges for all approaches to nonlinear unobserved effects models. CRE approaches for the case of lagged dependent variables, but with otherwise strictly exogenous covariates, are available for balanced panels; see Wooldridge (2000) for a summary. Wooldridge (2008) suggests an approach, under ignorable selection, that can work in the case of pure attrition (which imposes a particular pattern on the selection indicators). But more work needs to be done. For specific models, a balanced panel is not needed for certain methods that eliminate the heterogeneity – for example, Honoré and Kyriazidou (2000) for dynamic binary response, Honoré and Hu (2004) for dynamic corner solutions – but these methods have other restrictions and effectively require dropping lots of data. Fixed effects methods for large T, particularly with bias adjustments, seem promising, but their asymptotic properties with small T need not be good, and it is unclear how features such as time dummies and unit root processes can be handled.

Appendix. Proof of Proposition 2.1

The case $\theta_i = 1$ for all i is obvious, because then the estimate $\tilde{\boldsymbol{\beta}}$ is from the pooled regression $y_{it} - \bar{y}_i$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$ with $s_{it} = 1 -$ and this defines the FE estimate on the unbalanced panel. To handle other cases, we assume that the appropriate matrices are invertible. Generally, the invertibility requirement holds under standard assumptions of time-variation in the $\{\mathbf{x}_{it}\}$ and no perfect collinearity when $0 \leq \theta_i < 1$.

First consider the case without $\mathbf{z}_i$. Then $\tilde{\boldsymbol{\beta}}$ can be obtained from the Frisch–Waugh (partialling out) theorem for OLS. First, regress $\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i$ on $(1 - \theta_i)\bar{\mathbf{x}}_i$ (using the selected sample) and obtain the residuals, say $\tilde{\mathbf{r}}_{it}$. Then obtain $\tilde{\boldsymbol{\beta}}$ from the pooled OLS regression (again on the selected sample) of $y_{it} - \theta_i\bar{y}$ on $\tilde{\mathbf{r}}_{it}$. The residuals $\tilde{\mathbf{r}}_{it}$ are simple to obtain. We can write them as

$$\tilde{\mathbf{r}}_{it} = (\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i) - (1 - \theta_i)\bar{\mathbf{x}}_i\tilde{\boldsymbol{\Pi}}$$

where

$$\begin{aligned}\tilde{\boldsymbol{\Pi}} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)\ddot{\mathbf{x}}_i'(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)\ddot{\mathbf{x}}_i'\mathbf{x}_{it}-\sum_{i=1}^{N} T_i\theta_i(1-\theta_i)\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]^{-1}\left[\sum_{i=1}^{N} T_i(1-\theta_i)\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i-\sum_{i=1}^{N} T_i\theta_i(1-\theta_i)\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]^{-1}\left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right]=\mathbf{I}_K.\end{aligned}$$

It follows that $\tilde{\mathbf{r}}_{it}=(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)-(1-\theta_i)\bar{\mathbf{x}}_i=\mathbf{x}_{it}-\bar{\mathbf{x}}_i$, which is simply the time-demeaned covariates. Now we can write

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'(y_{it}-\theta_i\bar{y}_i)\right]\\ &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'y_{it}\right]\end{aligned}$$

using the fact $\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\theta_i\bar{y}_i=\theta_i\bar{y}_i\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'=\mathbf{0}$ because $\bar{\mathbf{x}}_i$ is the average over the selected time periods. But this final formula is just $\hat{\boldsymbol{\beta}}_{FE}$ on the selected sample.

For the case with $\mathbf{z}_i$, we can apply the Frisch–Waugh theorem again to obtain the appropriate residuals. That is, now $\tilde{\mathbf{r}}_{it}$ are from the regression $\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i$ on $(1-\theta_i)\bar{\mathbf{x}}_i$, $(1-\theta_i)\mathbf{z}_i$ with $s_{it}=1$. But now we partial out $\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i$ from $(1-\theta_i)\bar{\mathbf{x}}_i$ to get residuals $\tilde{\mathbf{q}}_{it}$, say, and we just showed $\tilde{\mathbf{q}}_{it}=\mathbf{x}_{it}-\bar{\mathbf{x}}_i$. The other residuals we need are from $(1-\theta_i)\mathbf{z}_i$ on $(1-\theta_i)\bar{\mathbf{x}}_i$ with $s_{it}=1$, and it is obvious that these, say $\tilde{\mathbf{e}}_i$, depend only on i. So the $\tilde{\mathbf{r}}_{it}$ are from $\mathbf{x}_{it}-\bar{\mathbf{x}}_i$ on $\tilde{\mathbf{e}}_i$ across i and t with $s_{it}=1$, and because $\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)=\mathbf{0}$ for all i, it follows that

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\mathbf{e}}_i'\tilde{\mathbf{q}}_{it}=\mathbf{0}.$$

This means $\tilde{\mathbf{r}}_{it}=\tilde{\mathbf{q}}_{it}=\mathbf{x}_{it}-\bar{\mathbf{x}}_i$, as before. Now the rest of the proof is the same. □

References

Abrevaya, J., 2013. The projection approach for unbalanced panels. Econom. J. 16, 161–178.

Altonji, J.G., Matzkin, R.L., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73, 1053–1102.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econom. Stud. 58, 277–297.

Baltagi, B., 2001. Econometric Analysis of Panel Data, second ed. Wiley, New York.

Blundell, R., Powell, J.L., 2003. Endogeneity in nonparametric and semiparametric regression models, with Richard Blundell. In: Dewatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, Vol. 2. Cambridge University Press, Cambridge, pp. 312–357.

Chamberlain, G., 1982. Multivariate regression models for panel data. J. Econometrics 1, 5–46.

Chamberlain, G., 1984. Panel data. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. 2. North Holland, Amsterdam, pp. 1248–1318.

Chernozhukov, V., Fernández-Val, I., Newey, W.K., 2013. Average and quantile effects in nonseparable panel models. Econometrica 81, 535–580.

Fernández-Val, I., 2009. Fixed effects estimation of structural parameters and marginal effects in panel probit models. J. Econometrics 150, 71–85.

Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel models with large N, T. J. Econometrics 192, 291–312.

Gourieroux, C.A., Monfort, A., Trognon, C., 1984. Pseudo-maximum likelihood methods: theory. Econometrica 52, 681–700.

Guggenberger, P., 2010. The impact of a Hausman pretest on the size of a hypothesis test: The panel data case. J. Econometrics 156, 337–343.

Hahn, J., Newey, W.K., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72, 1295–1319.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica.

Hayashi, F., 2001. Econometrics. Princeton University Press, Princeton, NJ.

Honoré, B.E., Hu, L., 2004. Estimation of cross sectional and panel data censored regression models with endogeneity. J. Econometrics 122, 293–316.

Honoré, B.E., Kyriazidou, E., 2000. Panel data discrete choice models with lagged dependent variables. Econometrica 68, 839–874.

Kwak, D.W., Martin, R., Wooldridge, J.M., 2018. The Robustness of the Fixed Effects Logit Estimator to Violations of Conditional Independence. Mimeo, Michigan State University Department of Economics.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.

Papke, L.E., Wooldridge, J.M., 2008. Panel data methods for fractional response variables with an application to test pass rates. J. Econometrics 145, 121–133.

Semykina, A., Wooldridge, J.M., 2010. Estimating panel data models in the presence of endogeneity and selection. J. Econometrics 157, 375–380.

Semykina, A., Wooldridge, J.M., 2017. Binary response panel data models with sample selection and self selection. J. Appl. Econometrics.

Verbeek, M., Nijman, T., 1996. Testing for selectivity bias in panel data. Internat. Econom. Rev. 33, 681–703.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.

Wooldridge, J.M., 1995. Selection corrections for panel data models under conditional mean independence assumptions. J. Econometrics 68, 115–132.
Wooldridge, J.M., 2000. A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. Econom. Lett. 68, 245–250.
Wooldridge, J.M., 2005. Unobserved heterogeneity and estimation of average partial effects. In: Andrews, D.W.K., Stock, J.H. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge, pp. 27–55.
Wooldridge, J.M., 2008. Nonlinear Dynamic Panel Data Models with Unobserved Effects, invited lecture, Canadian Econometrics Study Group, Montreal.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press, Cambridge, MA.