CORRELATED RANDOM EFFECTS MODELS WITH UNBALANCED PANELS

Jeffrey M. Wooldridge*
Department of Economics
Michigan State University
East Lansing, MI 48824-1038
wooldri1@msu.edu

This version: May 2010

*I presented an earlier version of this paper, called "Nonlinear Correlated Random Effects Models with Unbalanced Panels," at the 15th Conference on Panel Data, Bonn, Germany, July 3-5, 2009. I thank Simon Quinn for helpful comments.

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. Even for nonlinear models, in many cases the estimators can be implemented using standard software. The framework suggests straightforward tests of correlation between heterogeneity and the covariates, as well as sample selection that is correlation 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, Blundell and Powell (2003), Altonji and Matzkin (2005), and Wooldridge (2005)) shows that quantities of interest – usually called "average marginal effects" (AMEs) or "average partial effects" (APEs) – 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 (2002) 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.

The leading competitor to CRE approaches are so-called "fixed effects" (FE) methods, which, for the purposes of this paper, treat the heterogeneity as parameters to be estimated. (Perhaps a better characterization is that the FE approach studies the properties of the fixed population parameters and, more recently, average partial effects, when heterogeneity is handled via estimating separate parameters for each population unit.) As is well known, except in some very special cases, estimating unobserved heterogeneity for each unit in the sample generally suffer from the incidental parameters problem – both in estimating population parameters and APEs. Some headway has been made in obtaining bias-corrected versions of "fixed effects" estimators for nonlinear models – for example, Hahn and Newey (2004) and Fernandez-Val (2008). These methods are promising, but they currently have several practical shortcomings. First, the number of time periods needed for the bias adjustments to work well is often greater than is available in many applications. Second, an important point is that recent

bias adjustments include the assumptions of stationarity and weak dependence; in some cases, the very strong assumption of serial independence (conditional on the heterogeneity) is maintained. As a practical matter, sources of serial correlation in addition to that caused by unobserved heterogeneity is very common in empirical work. (For example, for linear models it is often the case that the idiosyncratic errors have strong forms of serial correlation quite apart from the correlation caused by the heterogeneity appearing in every time period.) The requirement of stationarity is also very strong and has substantive restrictions: it rules out staples in empirical work such as including separate year effects, which can be estimated very precisely given a large cross section. In addition, the technical problem of allowing separate period effects when large-sample approximations involve a growing number of time periods has not yet been solved and is likely to be difficult, as it effectively introduces an incidental parameters problem in the time series dimension to go along with that in the cross section dimension. As this literature currently stands, the restrictions on time series dependence are not just regularity conditions that simplify proofs; the adjustments themselves only make sense under stationarity and weak dependence. See Imbens and Wooldridge (2007) for a summary.

Recently, Chernozhukov, Fernández-Val, Hahn, and Newey (2009) (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, 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 (2003), in which case the APEs are point identified.) These methods are very promising but a still limited to discrete covariates. Plus, in some cases we may be willing to impose more restrictions in order to point identify the APEs.

Another method that is often used, but only in special cases, is conditional maximum likelihood estimation. The general approach is to find a conditional likelihood function that is free of the unobserved heterogeneity but depends on the population parameters. Unfortunately, this method has rather limited scope. It applies to linear models (but where we do not needed) and a few nonlinear models. It is most commonly applied to the unobserved effects logit model and also to the unobserved effects Poisson regression model. When it applies, the CMLE has the advantage that it puts no restrictions on the heterogeneity distribution – either unconditionally or conditionally. Unfortunately, even in the limited cases where it applies, CMLE can impose substantive restrictions. For example, the CMLE for the logit model is inconsistent if the conditional independence assumption fails – see Kwak and Wooldridge (2009). (Other CMLEs are more robust, such as those for the linear and Poisson unobserved effects models, but again these are special cases. See Wooldridge (1999) for the Poisson case.) A positive feature of the CMLE approach is that it works for any number of time periods and imposes no restrictions on the time series properties of the covariates. However, because CMLEs are intended to leave heterogeneity distributions unspecified, it is unclear how to obtain average partial effects. In other words, we cannot estimate the magnitudes of effects of covariates.

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 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, although restrictions such as exchangeability can be very useful – see Altonji and

Matzkin (2003). In other words, for estimation using balanced panels, CRE, FE, and CMLE involvate tradeoffs among assumptions and the type 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.

There is one clear disadvantage of CRE approaches when compared with either FE or CMLE methods: neither FE nor CMLE approaches require balanced panels whereas CRE methods, as currently developed, are for balanced panels. Generally, it is not obvious how to extend CRE approaches for balanced panels to unbalanced panels. In this paper I suggest an approach – which can be combined with recently proposed semiparametric and nonparametric methods if desired – and also provide simple implementations in the context of commonly used models, such as the CRE probit, ordered probit, and Tobit models.

A key assumption used in this paper is either implicit or explicit in most analyses with unbalanced panels, particularly when heterogeneity is removed or treated as parameters to estimate. Namely, sample selection is assumed not to be systematically related to unobserved shocks. (The exact statement of the assumption depends on whether the model is linear and where a full distribution has been specified or just a feature of it, such as a conditional mean. We state precise assumptions when appropriate.) Neverthleless, one of the attractions of, say, fixed effects estimation in the linear model – which we review briefly in Section 2 – is that selection can be arbitrarily correlated with unobserved hetergeneity. My approach to CRE models allows such correlation, too. In fact, the heterogeneity is allowed to be correlated with the entire history of selection and the (selected) covariates. Unlike CMLE approaches, I do not restrict the serial dependence in the data.

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. This result motivates the approaches in Sections 3 and 4 for more complicated models, but it is of interest in its own right because it leads to simple, fully robust Hausman specification tests for the unbalanced case. This section also discusses how one might test a subset of the exogeneity assumptions used by the usual RE estimator. 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 observable covariates in unbalanced panels.

Section 4 proposes a general method for allowing correlated random effects in nonlinear models. The motivation is given by the findings in Sections 2 and 3.

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 discusses how the methods can be applied to popular nonlinear models, such as probit (including for fractional variables), ordered probit, and Tobit. Simple tests for violation of the ignorability of selection are discussed in this section as well.

Section 7 contains a general proposal for comparing fit across different models. The approach appears to be new – whether or not we are studying a balanced panel – 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 summarize some results that are useful for testing key assumptions.

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

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, $s_{it} = 0$. This is very common in panel data applications with unbalanced panels.

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 $\mathbf{x}_{it}$ can generally include a fully set of time dummies, or other aggretate time varaibles. We view this as the equation that holds for underlying random variables in all T time periods. We are interested in this paper 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 $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iT})$. 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, note that 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 crurical for nonlinear models.

It is well-known – see, for example, Verbeek and Nijman (1996), Hayashi (2001) and Wooldridge (2002, 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." If selection in every time period is independent of the covariates and idiosyncratic errors in ever time period then we can get by with a zero correlation assumption between $\mathbf{x}_{ir}$ and u_{it} for all $r, t = 1, \ldots, T$ in the population. Because we are interested in nonlinear models, we will use assumptions stated in terms of

conditional means.

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

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

and when we average this equation across t for each i we get

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

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, is sufficient for consistency.

As a computational point that becomes more important in complicated models, 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 $(\mathbf{x}_{it}, y_{it})$. Consequently, there are often pairs (i,t) where we may observe some elements in $(\mathbf{x}_{it}, y_{it})$ but where the information on these varaibles is not used in estimation.

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, but adding the time averages of the covariates as additional explanatory variables. Perhaps less well known is that 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

full set of data on the covariates and response variables. Then, estimate 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. The coefficient vector $\hat{\boldsymbol{\beta}}$ is identical to the fixed effects (within) estimator on the unbalanced panel. Any aggregate time variables, including time dummies, should be part of $\mathbf{x}_{it}$, and their time averages must be included in $\bar{\mathbf{x}}_i$. The reason is that, unlike in the balanced case, the time average of aggregate time variables changes across i because we average different time periods for different i.

In addition, if we run any pooled regression of the form

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

where $\mathbf{z}_i$ is any vector of time-constant variables, then $\hat{\boldsymbol{\beta}}$ is still the fixed effects estimate. For example, if we add the number of time periods, T_i, or interactions of the form $T_i \cdot \bar{\mathbf{x}}_i$ (that is, the sums in addition to the averages), the estimated coefficients on $\mathbf{x}_{it}$ are the fixed effects estimates. The same is true if we allow a different set of coefficients on $\bar{\mathbf{x}}_i$ depending on T_i. Of course, we can also add variables such as gender in a wage equation.

As other examples of $\mathbf{z}_i$, one can use $(\bar{\mathbf{x}}_{i1}, T_{i1}, \bar{\mathbf{x}}_{i2}, T_{i2}, \ldots, \bar{\mathbf{x}}_{iG}, T_{iG})$ where we partition $\{1, 2, \ldots, T\}$ into G groups and then compute the averages of the selection observations, $\bar{\mathbf{x}}_{ig}$, and the total number of selected periods, T_{ig}. Because we can get $\bar{\mathbf{x}}_i$ as a linear combination of $(\bar{\mathbf{x}}_{i1}, \bar{\mathbf{x}}_{i2}, \ldots, \bar{\mathbf{x}}_{iG})$, the pooled OLS estimator of $\boldsymbol{\beta}$ is still the FE estimate. In other words, allowing for very general correlation between c_i and the (selected) sequence of covariates in the standard linear model produces an estimator that is commonly used, and is robust to any kind of correlation between c_i and $\{(\mathbf{x}_{i1}, s_{i1}), (\mathbf{x}_{i2}, s_{i2}), \ldots, (\mathbf{x}_{iT}, s_{it})\}$. When we discuss models with random slopes in the next section, and nonlinear models in Section 4, this point is useful

because we will have to take models relating heterogeneity to the covariates and selection more seriously. Yet at least in the leading case, the estimator is not sensitive to the specification of $E(c_i|\mathbf{x}_i, \mathbf{z}_i)$.

It is useful to have a general result that contains algebraic equivalances for pooled OLS as well as random effects. 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 $\mathbf{x}_{it}$ contains any aggregate time varaibles, the RE estimator can be obtained from the pooled OLS regression

$$y_{it} - \theta_i \bar{y}_i \text{ on } \mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i, \ (1 - \theta_i)\bar{\mathbf{x}}_i, \ (1 - \theta_i)\mathbf{z}_i \text{ if } s_{it} = 1, \tag{2.9}$$

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). (Of course, in practice, the variance parameters are replaced with estimates, but that is unimportant for an algebraic equivalance.) For our purposes, all that matters is that pooled OLS with the time averages added ($\theta_i = 0$) and random effects are special cases.

Proposition 2.1: Consider pooled OLS regressions of the form in (2.9), where the time averages are computed using the selected observations (so, for example, $\bar{\mathbf{x}}_i = T_i^{-1} \sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}$). Note that $\mathbf{z}_i$ can include the intercept and $\mathbf{x}_{it}$ any aggregate time variables. Let $\tilde{\boldsymbol{\beta}}$ be the vector $(K \times 1)$ 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.

Proof: 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 theorem. 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

$$\tilde{\mathbf{r}}_{it} = (\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i) - (1-\theta_i)\bar{\mathbf{x}}_i \tilde{\boldsymbol{\Pi}} \tag{2.10}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\Pi}} &= \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(1-\theta_i)^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right]^{-1} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(1-\theta_i) \bar{\mathbf{x}}_i' (\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i) \right] \\ &= \left[\sum_{i=1}^{N} T_i (1-\theta_i)^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right]^{-1} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(1-\theta_i) \bar{\mathbf{x}}_i' \mathbf{x}_{it} - \sum_{i=1}^{N} T_i \theta_i (1-\theta_i) \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i) \right] \\ &= \left[\sum_{i=1}^{N} T_i (1-\theta_i)^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right]^{-1} \left[\sum_{i=1}^{N} T_i (1-\theta_i) \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i - \sum_{i=1}^{N} T_i \theta_i (1-\theta_i) \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i) \right] \\ &= \left[\sum_{i=1}^{N} T_i (1-\theta_i)^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right]^{-1} \left[\sum_{i=1}^{N} T_i (1-\theta_i)^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right] = \mathbf{I}_K. \end{aligned} \tag{2.11}$$

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

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right]^{-1} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \theta_i \bar{y}_i) \right] \\ &= \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right]^{-1} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' y_{it} \right] \end{aligned} \tag{2.12}$$

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. ■

The conclusions of Proposition 2.1 verify the previous claims made for pooled OLS as well as random effects on the unbalanced panel. A nice application of this algebraic equivalance result is a simple way to obtain regression-based, fully robust Hausman tests using unbalanced panels. Write a model with time-constant variables $\mathbf{z}_i$ as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + c_i + u_{it},\ t = 1, \ldots, T, \tag{2.13}$$

where, again, we use a data point if $s_{it} = 1$. Assume that $\mathbf{z}_i$ includes a constant. If we use the Mundlak equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \mathbf{z}_i\boldsymbol{\gamma} + a_i + u_{it} \tag{2.14}$$

and estimate this by, say, RE, we know from Proposition 2.1 that the estimate of $\boldsymbol{\beta}$ is the FE estimate. Now the regression based Hausman test is just, say, a Wald test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$ after RE estimation of the augmented equation. Any unit with $T_i = 1$ can be included in the estimation and testing, but, of course, if we only have units with $T_i = 1$ then $\bar{\mathbf{x}}_i = \mathbf{x}_{it}$ for the single time period t with $s_{it} = 1$; thus, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ cannot be distinguished.

We can also use Mundlak's CRE formulation, whether the panel is balanced or not, to test

a subset of coefficients in $\boldsymbol{\xi}$. For example, we might postulate

$$E(c_i|\mathbf{x}_i, \mathbf{z}_i) = E(c_i|\bar{\mathbf{x}}_i, \mathbf{z}_i) = E(c_i|\bar{\mathbf{x}}_{i(1)}, \mathbf{z}_i), \tag{2.15}$$

where $\bar{\mathbf{x}}_{i(1)}$ is $\bar{\mathbf{x}}_i$ but with the first element, $\bar{x}_{i1}$, removed. This allows us to test the possibility that, after controlling for $(\bar{x}_{i2}, \ldots, \bar{x}_{iK}, \mathbf{z}_i)$, $\{x_{it1}\}$ is exogenous with respect to c_i (as well as $\{u_{it}\}$). The test is just a fully robust t test of $H_0 : \xi_1 = 0$. A failure to reject provides some justification for estimating the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \xi_2\bar{x}_{i2} + \ldots + \xi_K\bar{x}_{iK} + \mathbf{z}_i\boldsymbol{\gamma} + c_i + u_{it} \tag{2.16}$$

by random effects (on the unbalanced panel) – likely leading to a more precise, perhaps much more precise, estimate of β_1 (the coefficient on x_{it1}). Naturally, one should make inference fully robust to heteroskedasticity in the composite error and serial correlation in $\{u_{it}\}$.

Using a t statistic on $\hat{\xi}_1$ is not the same, even asymptotically, as comparing $\hat{\beta}_{FE,1}$ and $\hat{\beta}_{RE,1}$ via a one degree-of-freedom Hausman test. The latter maintains $\boldsymbol{\xi} = \mathbf{0}$ under the null – that is, the RE estimator of β_1 is generally inconsistent under (2.15) – whereas a t test of $H_0 : \xi_1 = 0$ is silent on other elements of $\boldsymbol{\xi}$. If β_1 is the coefficient of primary interest, it may make more sense to test $H_0 : \xi_1 = 0$, as it allows (partial) correlation between c_i and $\bar{\mathbf{x}}_{i(1)}$.

Excluding $\bar{x}_{i1}$ from (2.16) is in the spirit of imposing extra restrictions to estimate the parameters in (2.13). Hausman and Taylor (1981) use exogeneity assumptions on both time-constant and time-varying covariates in (2.13), mainly to identify elements of $\boldsymbol{\gamma}$ when the full RE orthogonality conditions do not hold. In (2.16), we do not need to exclude $\bar{x}_{i1}$ in order to identify β_1 (because $\{x_{it1}\}$ has some time variation), but doing so may result in an estimate of β_1 with more precision.

In summary, this section has shown that even if we make a very strong assumption in the

unbalanced case, namely, $c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$, $E(a_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}) = 0$, the resulting estimator – pooled OLS or RE – is identical to an estimator, fixed effects, that 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 for the balanced case suggests that in models with more complicated functional forms and heterogeneity, simple models of $E(c_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\})$ may work reasonably well.

3. Linear Models with Correlated Random Slopes

If we start with a model that has individual-specific slopes, the presence of unbalanced panels is more difficult to treat. Wooldridge (2005) shows that using fixed effects in a linear model where the random slopes are ignored has some robustness properties for estimating the average effect. But those findings do not carry over to unbalanced panels where selection may be correlated with heterogeneity because the slope heterogeneity becomes part of the error term.

To see how to handle 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 also assume 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

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

This is an obvious extension of assumption (2.3).

In what follows, we assume that we do not want to select elements of $\mathbf{b}_i$ that are allowed to change with *i*. 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 (2002, Chapter 11). This approach is attractive in specific instances, but it cannot be used in general (and not at all for nonlinear models).

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

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

Now, because we only use observations with $s_{it} = 1$, 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. That is, we condition on $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. 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 then if 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. We want to be able to handle cases where the covariates are missing, too, as happens when units are simply not observed at all for some time periods.

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

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

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.

At this point 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 implicity 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 $E(s_i y_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta}$. Pooled estimation using the selected sample or GLS 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 exchangeable functions of $\{\mathbf{h}_{it} : t = 1, \ldots, T\}$. In the balanced panel case, this approach was suggested by Altonji and Matzkin (2005) in a fully nonparametric setting. The leading examples of exchangeable functions are sums (or 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 the exchangeable functions satisfying

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

Further, if we assume that these expectations depend only on $\bar{\mathbf{x}}_i$, and in a linear fashion, we have

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

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

where $\boldsymbol{\mu}_{\bar{\mathbf{x}}_i} = E(\bar{\mathbf{x}}_i)$ is subtracted from $\bar{\mathbf{x}}_i$ to ensure the zero unconditional means of c_i and $\mathbf{d}_i$. If we insert these expectations into (3.6) we obtain

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

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 (adjusted, of course, for the unbalanced panel), 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, and that selection might be correlated with the 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 (3.14) 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

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

and then allow these to be correlated with c_i and $\mathbf{d}_i$. We omit the details so that we can move on to nonlinear models.

As a final comment before we turn to nonlinear models, note that in any of the previous formulations we have simple tests of dynamic selection bias available. We have assumed that our model for $E(c_i, \mathbf{d}_i | \{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ captures how heterogeneity depends on entire sequence of selection and the sequence of selected covariates. Therefore, under the ignorability assumption (3.3), no other functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ should appear in $E(s_i y_{it} | \mathbf{h}_i)$. Of course, we cannot include s_{it} as an explanatory variable at time t because we only use data with $t = 1$. But we can use lagged and lead values. A simple and possibly revealing test is to add as extra regressors at time t the variables $(s_{i,t+1}, s_{i,t+1}\mathbf{x}_{i,t+1})$. We can compute a fully robust (to serial correlation and heteroskedasticity) Wald test of the null hypothesis that (3.3) holds and that our model for $E(c_i, \mathbf{d}_i | \mathbf{h}_i)$ is correct as the joint significance test of $(s_{i,t+1}, s_{i,t+1}\mathbf{x}_{i,t+1})$. (If we want more of a pure test for selection bias, we can include just $s_{i,t+1}$ and just use a robust t

statistic.) If we reject the null we might have a selection problem where being in the sample at time $t + 1$ is correlated with shocks to y at time t, that is, $s_{i,t+1}$ is correlated with u_{it}. (In carrying out this test, our hope is that the time-constant parts of the error term are uncorrelated with the entire history, $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$, as is the case when we have properly modeled $E(c_i, \mathbf{d}_i|\mathbf{h}_i)$.)

4. A Modeling Approach for Nonlinear Models

We can apply the general approach for linear models with random slopes to general nonlinear models, although in some cases we have to work in terms of conditional distributions rather than conditional means. We assume that interest lies in the distribution

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

where, in generaly, $\mathbf{y}_{it}$ can be a vector and $\mathbf{x}_{it}$ is a set of observed conditioning variables. In this section,we denote the vector of heterogeneity by $\mathbf{c}_i$. We also restrict attention in this paper to strictly exogenous covariates, so that we impose the substantive restriction

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

where, again, $\mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}\}$ is the entire history of covariates (whether or not the entire history is observed). We assume that we have specified, for each t, a density for $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, which we write with dummy 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 the (conditional) independence over time. Plus, as discussed in Wooldridge (2002), the 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 conditonal 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, this assumption allows selection to be abitrarily correlated with $(\mathbf{x}_i, \mathbf{c}_i)$ but not generally with "shocks" to $\mathbf{y}_{it}$.

As in the case of the linear model, our correlated random effects approach will be to specify 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

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

just as 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: that 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 (at least in principle) 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 (abusing notation by not distinguishing the true parameters from the dummy arguments) is

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

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

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})], \tag{4.9}$$

and in leading cases – as we will see in Section 7 – the partial log likelihood for the unbalanced case is simple to compute. Further, the large-*N*, fixed-*T* asymptotics is straightforward. We do not provide regularity conditions here because in most CRE applications the log likelihoods are very smooth.

In general, inference needs to be made robust to the serial dependence in the scores from (4.9). Let $\boldsymbol{\theta}$ be the vector of all parameters, and assume identification holds along with regularity conditions. Further, 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

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

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

Notice that the definition of **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 $\hat{\boldsymbol{\theta}}$. In many cases we can replace $\mathbf{H}_{it}(\boldsymbol{\theta})$ with its expectation conditional on $\mathbf{w}_i$.

Before we turn to estimating quantities of interest, it is useful to know that essentially the same arguments carry through for estimating conditional means. 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 $\mathbf{w}_i$ (which, again, is valid for the mean when $s_{it} = 1$). Of course, 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. We will cover some examples in Section 6.

5. Estimating Average Partial Effects

In most nonlinear models, the parameters $\boldsymbol{\gamma}$ appearing in $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$ provide only part of the story for 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,

generally for the setup described in Section 4 we have enough information to identify and estimate partial effects with the hetereogeneity 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 Imbens and Wooldridge (2007), 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) are 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 (2002) 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 deta 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.

As a general approach to flexibly estimating APEs, we might choose to estimate a separate model for every possible value of T_i (except $T_i = 1$, which is ruled out by most standard models and choices of $D(\mathbf{c}_i|\mathbf{w}_i)$). Suppose, for example, that we choose $\mathbf{w}_i = (T_i, \bar{\mathbf{x}}_i)$. In practice, the structure of the density $f_t(\mathbf{y}_t|\mathbf{x}_{it}, T_i, \bar{\mathbf{x}}_i; \boldsymbol{\theta})$ is the same across all values of T_i. (We will see examples in the next section.) Thus, for $r = 2, \ldots, T$, we estimate a vector of parameters, $\hat{\boldsymbol{\theta}}_r$, using pooled MLE or QMLE with $T_i = r$. We can then estimate the ASF for each t as

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\sum_{r=2}^{T} 1[T_i = r]q_t(\mathbf{x}_t, \bar{\mathbf{x}}_i; \hat{\boldsymbol{\theta}}_r) \tag{5.5}$$

where $q_t(\mathbf{x}_t, \bar{\mathbf{x}}_i; \hat{\boldsymbol{\theta}}_r)$ is the estimated conditional mean function using the $T_i = r$ observations. Of course, because (5.5) does not include the $T_i = 1$ observations, the estimated APEs necessarily exclude that part of the population. Perhaps this is as it should be because, just as we saw in Section 4 for the linear model with random coefficients, the $T_i = 1$ observations cannot be used to estimate the coefficients. In some cases, though, we may wish to impose enough constant coefficients in $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ so that the $T_i = 1$ observations are helpful and used in estimation.

With an unbalanced panel, there is a somewhat subtle point about computing a single

average partial effect from the ASF. With a balanced panel, it is natural to average $\widehat{APE}_{tj}(\mathbf{x}_t)$ across the distribution of $\mathbf{x}_{it}$, and then possibly across t, too. With a random sample in each time period, this averaging is straightforward. But if selection s_{it} depends on $\mathbf{x}_{it}$, averaging across the selected sample does not consistently estimate $E_{\mathbf{x}_{it}}[APE_{tj}(\mathbf{x}_{it})]$. Presumably we still have an idea of useful values to plug in for $\mathbf{x}_t$, but generally estimating specific features of the distribution of $\mathbf{x}_{it}$ can be difficult. We might have to be satisfied with computing the average of $APE_{tj}(\mathbf{x}_{it})$ for the selected sample, that is, $E[s_{it}APE_{tj}(\mathbf{x}_{it})]$.

6. Examples

We now consider a few examples to show how simply the proposed methods apply to standard models. We begin with the unobserved effects probit model without restricting serial dependence. 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. 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 imposing 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 orginal 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, Monfort, and Trognon (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 heteroskedastict 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 spirt of allowing random slopes on $\mathbf{x}_{it}$ in the orginal 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 $E(y_{it}|\mathbf{x}_{it}, c_i)$. We could more formally take a semiparametric approach and assume, say, that $D(c_i|\mathbf{w}_i)$ depends on a linear index in $(1[T_i = 2], \ldots, 1[T_i = T], 1[T_i = 2] \cdot \bar{\mathbf{x}}_i, \ldots, 1[T_i = T] \cdot \bar{\mathbf{x}}_i)$. Then, we can modify the Blundell and Powell (2003) approach for cross section data with endogenous explanatory variables for the current panel data setting.

Other approaches to estimation are possible. For example, the key ignorability of selection assumption justifies estimation on any balanced subset of data. So we could only use observations with $T_i = T$, say, and then apply the usual CRE methods for the balanced case;

see Wooldridge (2002) and Imbens and Wooldridge (2007). This is attractive in situations where the vast majority of observations have a full set of time periods. (Technically, we replace the selection indicator, s_{it}, in (4.9) with the product, $s_{i1}s_{i2}\cdots s_{iT}$ to pick out observations with a full set of time periods.) We can also pick out, say, pairs of observations; and so on.

An estimation approach that may be more efficient than just pooling is minimum distance estimation. In the current setting, we can estimate a different set of parameters, including for $\boldsymbol{\beta}$, for each $T_i = 2, \ldots, T$, and then impose the restrictions across r using minimum distance. A less efficient but more flexible approach would be to estimate a standard CRE probit model for each $T_i = 2, \ldots, T$ (allowing the variance to change only with T_i and not $\bar{\mathbf{x}}_i$). This would give us (implicitly scaled) coefficients $(\hat{\boldsymbol{\beta}}_r, \hat{\psi}_r, \hat{\boldsymbol{\xi}}_r)$ for each r. We can easily compute the ASFs conditional on each r as

$$\widehat{ASF}_r(\mathbf{x}_t) = N_r^{-1} \sum_{i=1}^{N} 1[T_i = r]\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_r + \hat{\psi}_r + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r), \tag{6.10}$$

where N_r is the number of i with $T_i = r$. Thes weighted average of these across r is an estimate of the overall ASF for each t (again, as a function of $\mathbf{x}_t$):

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} \sum_{r=2}^{T} 1[T_i = r]\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_r + \hat{\psi}_r + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r), \tag{6.11}$$

Allowing heteroskedasticity as a function of $\bar{\mathbf{x}}_i$ is almost trivial because for each T_i we can use heteroskedastic probit to estimate the coefficients. Then

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} \sum_{r=2}^{T} 1[T_i = r]\Phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}}_r + \hat{\psi}_r + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r}{\exp(\bar{\mathbf{x}}_i\hat{\boldsymbol{\lambda}}_r/2)}\right], \tag{6.12}$$

where $\hat{\boldsymbol{\lambda}}_r$ is the vector of variance parameters. Even though estimates for each r may not be

especially precise,.averaging across all of the estimates can lead to precise estimates of the ASF. For even more flexibility we can add interactions $\bar{\mathbf{x}}_i \otimes \mathbf{x}_{it}$ in the estimation for each r, with or without heteroskedasticity. In the more general case, the ASF is then estimated as

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} \sum_{r=2}^{T} 1[T_i = r] \Phi \left[\frac{\mathbf{x}_t \hat{\boldsymbol{\beta}}_r + \hat{\psi}_r + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_r + (\bar{\mathbf{x}}_i \otimes \mathbf{x}_t) \hat{\boldsymbol{\eta}}_r}{\exp(\bar{\mathbf{x}}_i \hat{\boldsymbol{\lambda}}_r / 2)} \right], \tag{6.13}$$

so that partial effects with respect to the elements of $\mathbf{x}_t$ need to account for the interactions with the time averages, $\bar{\mathbf{x}}_i$. The specification underlying (6.13) is very quite flexible, but it does mean the $T_i = 1$ observations are dropped for computing the APEs. If we assumed independent, identically distributed $(s_{it}, \mathbf{x}_{it})$, $E(c_i | T_i, \bar{\mathbf{x}}_i) = T_i \bar{\mathbf{x}}_i \boldsymbol{\xi}$ and $Var(c_i | T_i, \bar{\mathbf{x}}_i) = \exp[\omega_0 + \omega_1 \log(T_i)]$. Thus, we could use interactions with T_i and $\bar{\mathbf{x}}_i$ in the "mean" part of the probit and simply $\log(T_i)$ in the variance part, possibly also adding $\bar{\mathbf{x}}_i$ for additional flexibility. Of course, this allows us to include the $T_i = 1$ observations at the cost of more assumptions.

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 sample variances and covariances, individual-specific trends, or break the time period into intervals and use averages over those intervals.

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 previous estimations and conducting a robust, joint test.

Everything just covered for the probit (or fractional probit) case extends to, say, the ordered probit case. Further, one can use some new strategies for handling multinomial responses that are computationally simple. 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 – in this case, for the response probabilities – are obtained by averaging out $\mathbf{w}_i$ when it is all done. 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. A Proposal for Goodness of Fit

An issue that arises in comparing across 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 a fractional response model, and we want to compare a linear model – with just a single, additive heterogeneity – to a fractional response model, also with a single source of hetereogeneity. If the linear model is estimated by fixed effects and the fractional model using 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.

By recognizing that the linear fixed effects estimator is actually a CRE estimator, we can consider the goodness-of-fit problem in a unified setting. 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 $D(\mathbf{c}_i|\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots,\mathbf{x}_{iT})=D(\mathbf{c}_i|\mathbf{w}_i)$. 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 the 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,

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

The density $f_t(\mathbf{y}_t|\mathbf{x}_t,\mathbf{w};\boldsymbol{\gamma},\boldsymbol{\delta})$ is obtained from the densities $g_t(\mathbf{y}_t|\mathbf{x}_t,\mathbf{c};\boldsymbol{\gamma})$ and $h(\mathbf{c}|\mathbf{w};\boldsymbol{\delta})$ – including the choice of $\mathbf{w}$ – and so the goodness-of-fit based on the log likelihood depends on the choice of $g_t(\mathbf{y}_t|\mathbf{x}_t,\mathbf{c};\boldsymbol{\gamma})$ and $h(\mathbf{c}|\mathbf{w};\boldsymbol{\delta})$, including which functions of $\{\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}\}$ are allowed to be related to $\mathbf{c}_i$.Different choices can be compared. Naturally, we can add penalties to the log likelihood for the number of overall parameters, as is done in with the Bayesian and Akaike information criteria.

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

In many cases we do not want to specify an entire conditional density $f_t(\mathbf{y}_t|\mathbf{x}_t,\mathbf{w};\boldsymbol{\gamma},\boldsymbol{\delta})$. Even if we do, we might be directly interested in the fit of the mean – something we can compare across models where we may or may not specify a full conditional distribution. As in Section 5, let $q_t(\mathbf{x}_{it},\mathbf{w}_i)$ be $E(y_{it}|\mathbf{x}_{it},\mathbf{w}_i,s_{it} = 1)$. Then we can compute a sum of squared residuals on the unbalanced panel (with a balanced panel being as special case) as

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

Again, this measure of fit is comparable across models that differ in $E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$ and $D(\mathbf{c}_i|\mathbf{w}_i)$, including the choice of $\mathbf{w}_i$. We can compare, say, a linear model estimated by fixed effects to a CRE fractional response model by using

$$q_t(\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}) = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \sum_{r=2}^{T-1} 1[T_i = r]\omega_r$$

for the linear case and, say,

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

for the nonlinear case. We can again use penalties for number of parameters if desired.

8. Concluding Remarks

I have offered some simple strategies for allowing unbalanced panels in correlated random effects models. Hopefully these methods make applying CRE models to panel data sets collected in practice somewhat easier. The nature of the approach – which extends the balanced case – is the need to model $D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ in terms of a set of sufficient statistics, $\mathbf{w}_i$. I have focused on parametric approximations, but the general approaches of Altonji and Matzkin (2005) and Blundell and Powell (2003, 2004)

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, the restrictions on the covariate (and, in this case, the selection process) such that $D(\mathbf{c}_i|\mathbf{w}_i)$ is normal for any T are quite strong. Technically, this is a valid criticism, and it motivates pursuing the nonparametric and semiparametric approaches cited above. Still, 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.

One we focus on average partial effects, we can think of the original specifications $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ and $D(\mathbf{c}_i|\mathbf{w}_i)$ as convenient ways to obtain estimable distributions $D(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$. We can choose, say, $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ in ways that are more flexible than allowed by either fixed effects approaches that treat $\mathbf{c}_i$ as parameters to estimate or conditional MLE approaches that try to eliminate $\mathbf{c}_i$. For example, of we start with a probit model $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$, 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 – while computationally intensive if we carry through with the random slopes probit model – is tractable, and has known large-N properties if we properly model $D(\mathbf{c}_i|\mathbf{w}_i)$. Moreoever, 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 requires a fairly sophisticated simulation study.

I have focused on pooled estimation methods. These are simple but may be inefficient. I mentioned the possibility of minimum distance estimation when we have restrictions imposed across the different values of T_i. But other possibilities suggest themselves. For example, we might restrict the joint distribution $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{x}_i, \mathbf{c}_i)$, as is common in pure random effects or CRE approaches with balanced panel. (Usually independence is assumed conditional on $(\mathbf{x}_i, \mathbf{c}_i)$.) Of course, such methods are generally more difficult computationally, but the general approach should carry through with the unconfoundedness assumption $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i) = D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i)$.

Another possibility, which is more directly applicable, is to use a generalized estimating equation (GEE) approach. GEE is effectively a multivariate weighted nonlinear least squares method, where the marginal means are assumed to be correctly specified but the joint distribution is unrestricted – as in the pooled case. But GEE attempts to exploit the neglected correlation over time by specifying simple correlation patters. In the case of a binary or fractional response, specifications such as (6.6) with the $\omega_r = 0$ are straightforward to handle using GEE software. See Imbens and Wooldridge (2007) or Papke and Wooldridge (2008) for further discussion with balanced panels.

The assumption of strictly exogenous covariates is strong and needs to be relaxed. Of

course, 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 (2005) for a summary. Wooldridge (2008) suggests an approach, under ignorable selection, that can work in the case of pure attrition (which imposes a particular patter 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.

References

Altonji, J.G. and R.L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica* 73, 1053-1102.

Baltagi, B. (2001), *Econometric Analysis of Panel Data*, 2e. Wiley: New York.

Blundell, R. and J.L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models, with Richard Blundell," in *Advances in Economics and Econonometrics: Theory and Applications*, Eighth World Congress, Volume 2, M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds. Cambridge: Cambridge University Press, 312-357.

Blundell, R. and J.L. Powell (2004), "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71, 655-679.

Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics* 1, 5-46.

Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics*, Volume 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North Holland, 1248-1318.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W.K. Newey (2009), "Identification and Estimation of Marginal Effects in Nonlinear Panel Models." Mimeo, Boston University Department of Economics.

Fernández-Val, I. (2008), "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models." Mimeo, Boston University Department of Economics.

Hahn, J. and W.K. Newey (2004), "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, 1295-1319.

Hayashi, F. (2001), *Econometrics*. Princeton University Press: Princeton, NJ.

Honoré, B.E. and L. Hu (2004), "Estimation of Cross Sectional and Panel Data Censored

Regression Models with Endogeneity," *Journal of Econometrics* 122, 293-316.

Honoré, B.E. and E. Kyriazidou (2000), "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica* 68, 839-874.

Imbens, G.W. and J.M. Wooldridge (2007), "What's New in Econometrics?" NBER Research Summer Institute, Cambridge, July/August, 2007.

Kwak, D.W. and J.M. Wooldridge (2009), "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. and J.M. Wooldridge (2008), "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," *Journal of Econometrics* 145, 121-133.

Verbeerk, M. and T. Nijman (1996), "Testing for Selectivity Bias in Panel Data," *International Economic Review* 33, 681-703.

Wooldridge, J.M. (1999), "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90, 77-97.

Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.

Wooldridge, J.M. (2005), "Unobserved Heterogeneity and Estimation of Average Partial Effects," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D.W.K. Andrews and J.H. Stock. Cambridge: Cambridge University Press, 27-55.

Wooldridge, J.M. (2008), "Nonlinear Dynamic Panel Data Models with Unobserved

Effects," invited lecture, Canadian Econometrics Study Group, Montreal, September 2008.