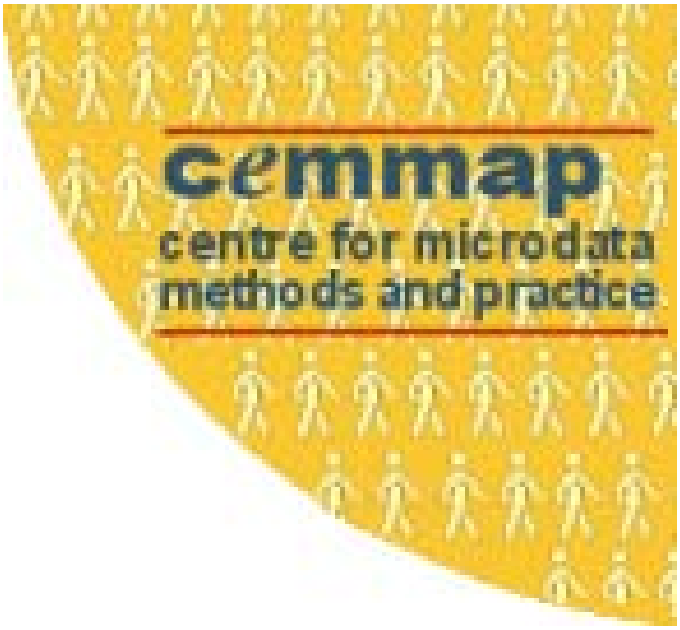

On The Robustness Of Fixed Effects And Related Estimators In Correlated Random Coefficient Panel Data Models

Jeffrey M. Wooldridge

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP04/04

ON THE ROBUSTNESS OF FIXED EFFECTS AND RELATED ESTIMATORS IN CORRELATED RANDOM COEFFICIENT PANEL DATA MODELS

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

This version: June 2003

Abstract: I show that a class of fixed effects estimators is reasonably robust for estimating the population-averaged slope coefficients in panel data models with individual-specific slopes, where the slopes are allowed to be correlated with the covariates. In addition to including the usual fixed effects estimator, the results apply to estimators that eliminate individual-specific trends. Further, asymptotic variance matrices are straightforward to estimate. I apply the results, and propose alternative estimators, to estimation of average treatment in a general class of unobserved effects models.

Keywords: Average Treatment Effect, Fixed Effects, Random Coefficient

*Thanks to Josh Angrist for reminding me about the example in Jin Hahn's JBES comment.

1. Introduction

The standard fixed effects, or within, estimator is a workhorse in empirical studies that rely on linear panel data models. When the partial effects of interest are on time-varying covariates, fixed effects estimation is attractive because it allows for additive, unobserved heterogeneity that can be arbitrarily correlated with the time-varying covariates. (On the other hand, with random effects methods we assume that unobserved heterogeneity is uncorrelated with observed covariates.) Extensions of the standard linear model with an additive unobserved effect include the random trend model, where each cross-sectional unit is allowed to have its own linear trend (in addition to a separate level effect); a special case is the so-called random growth model, as in Heckman and Hotz (1989). Wooldridge (2002a, Section 11.2) provides an overview of these kinds of models.

The properties of fixed effects estimators in general unobserved effects models have been derived assuming constant coefficients on the individual-specific, time-varying covariates. In Wooldridge (2003), I pointed out that the usual fixed effects estimator in the standard additive model is consistent in a model with individual specific slopes whenever the slopes are conditionally mean independent of the time-demeaned covariates. Importantly, this finding implies that the individual-specific slopes can be correlated with the time averages of the covariates, which we tend to think of as the major source of endogeneity in random coefficient panel data models. [With a small number of time periods, much more has been written about random coefficient models when the coefficients are assumed to be independent of the

covariates. See, for example, Hsiao (1986, Chapter 6). For most economic applications, the independence assumption is unrealistic.]

In this paper, I extend the framework of Wooldridge (2003) to allow for general aggregate time effects. I show that the fixed effects estimator that sweeps away the individual-specific intercept and slopes on the aggregate variables is satisfyingly robust to the presence of individual-specific slopes on the individual-specific covariates. Section 2 contains the main result. In Section 3, I briefly consider estimation strategies based on first differencing. As a special case in Section 4, I treat the so-called "random trend" model, which has become popular in empirical studies [for example, Papke (1994), Hoxby (1996), and Friedberg (1998)]. In addition, I cover average treatment effect estimation in a general unobserved effects model, generalizing a simple example due to Hahn (2001). Because consistency of the fixed effects estimator generally rules out aggregate time effects in index models, I also propose modified estimators that can consistently estimate the average treatment effects when fixed effects does not. In Section 5, I show how the result extends to models where time-constant observable covariates are available and correlated with unobserved heterogeneity.

2. Linear Models and a Result for Fixed Effects

For a random draw i from the population, the model is

$$y_{it} = \mathbf{w}_t \mathbf{a}_i + \mathbf{x}_{it} \mathbf{b}_i + u_{it}, t = 1, \ldots, T \tag{2.1}$$

where $\mathbf{w}_t$ is a $1 \times J$ vector of aggregate time variables – which we treat as nonrandom (without

consequence, since they are usually just time trends) – $\mathbf{a}_i$ is a $J \times 1$ vector of individual-specific slopes on the aggregate variables, $\mathbf{x}_{it}$ is a $1 \times K$ vector of covariates that change across time (possibly including year dummies), $\mathbf{b}_i$ is a $K \times 1$ vector of individual-specific slopes, and u_{it} is an idiosyncratic error. In what follows, we view T as being relatively small, and so we keep it as fixed in the asymptotic analysis. We assume we have a sample of size N randomly drawn from the population. For simplicity, we assume a balanced panel.

The object of interest is $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$, the $K \times 1$ vector of population-averaged partial effects. With small T, it is not possible to get precise estimates of each $\mathbf{b}_i$ (when we treat them as parameters to estimate). Instead, we hope to estimate the average effects using standard estimators. Throughout we maintain the assumption

$$\mathrm{E}(u_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{a}_i, \mathbf{b}_i) = 0, t = 1, \ldots, T, \tag{2.2}$$

which follows under the conditional mean assumption

$$\mathrm{E}(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{a}_i, \mathbf{b}_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it}, \mathbf{a}_i, \mathbf{b}_i) = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i, t = 1, \ldots, T. \tag{2.3}$$

Assumption (2.3) is a standard strict exogeneity condition in unobserved effects models: conditional on $(\mathbf{x}_{it}, a_i, b_i)$, the covariates from the other time periods do not affect the expected value of y_{it}. While this rules out the possibility of lagged dependent variables, it does not restrict the correlation between $(\mathbf{a}_i, \mathbf{b}_i)$ and $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$. The possibility of correlation between $\mathbf{b}_i$ and the $\mathbf{x}_{it}$ makes (2.1) a *correlated* random coefficient model, to borrow a phrase from Heckman and Vytlacil (1998) for the cross-sectional case.

The basic unobserved effects model is obtained with $w_t \equiv 1$ and $\mathbf{b}_i = \boldsymbol{\beta}$. The random linear trend model also has $\mathbf{b}_i = \beta$ but $\mathbf{w}_t \equiv (1, t)$, so that $\mathbf{a}_i = (a_{i1}, a_{i2})$, where a_{i2} is the

random trend for unit i. More flexible trends can be allowed with a sufficient number of time periods; for example, we can take $\mathbf{w}_t \equiv (1, t, t^2)$. However, we cannot allow a full set of year dummies to interact with separate unobserved heterogeneity terms, since we then lose identification of $\boldsymbol{\beta}$. Generally, we must have $J < T$; see, for example, Wooldridge (2002a, Section 11.2).

One possibility for analyzing equation (2.1) is to treat the $\mathbf{a}_i$ and $\mathbf{b}_i$ as parameters to estimate for for i. Under (2.2) and an appropriate rank condition, we can obtain unbiased estimators of $\mathbf{a}_i$ and $\mathbf{b}_i$, say $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{b}}_i$, by using ordinary least squares (OLS) on the time series for each i. Unfortunately, when T is small, the scope of such a strategy is limited. For one, we would need $J + K \leq T$ to even implement the procedure. Unless T is fairly large, precise estimation of $\mathbf{b}_i$ is not possible. Nevertheless, the average of the $\hat{\mathbf{b}}_i$ is generally consistent for $\boldsymbol{\beta}$ (for fixed T as $N \to \infty$) and $\sqrt{N}$-asymptotically normal. [See Wooldridge (2002a, Section 11.2) for verification in a closely related context.] Since this strategy is not available for large K, and since the covariance matrix of the resulting estimator is not easy to estimate, alternative methods for estimating the average effect are desirable.

In this paper, we study estimators of $\boldsymbol{\beta}$ that are *motivated* by the assumption that the slopes $\mathbf{b}_i$ are constant, but we study the properties of these estimators in the context of model (2.1). Write $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$, where $\mathrm{E}(\mathbf{d}_i) = 0$ by definition. Simple substitution into (2.1) gives

$$y_{it} = \mathbf{w}_t \mathbf{a}_i + \mathbf{x}_{it}\boldsymbol{\beta} + (\mathbf{x}_{it}\mathbf{d}_i + u_{it}) \tag{2.4}$$

$$\equiv \mathbf{w}_t \mathbf{a}_i + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \tag{2.5}$$

where $v_{it} \equiv \mathbf{x}_{it}\mathbf{d}_i + u_{it}$. Whether any or all of the elements of $\mathbf{a}_i$ are constant, we estimate $\boldsymbol{\beta}$ in (2.1) allowing the entire vector $\mathbf{a}_i$ to vary by i, and to be arbitrarily correlated with the $\mathbf{x}_{it}$. For the linear, additive effects model, this leads to the usual fixed effects estimator. More

generally, define $\mathbf{y}_i$ to be the $T \times 1$ vector of y_{it}, let $\mathbf{W}$ be the $T \times J$ matrix with t^{th} row $\mathbf{w}_t$, let $\mathbf{X}_i$ be the $T \times K$ matrix with t^{th} row $\mathbf{x}_{it}$, and let $\mathbf{v}_i$ be the vector of v_{it}. Then we can write

$$\mathbf{y}_i = \mathbf{W}\mathbf{a}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i = \mathbf{W}\mathbf{a}_i + \mathbf{X}_i\boldsymbol{\beta} + (\mathbf{X}_i\mathbf{d}_i + \mathbf{u}_i). \tag{2.6}$$

To eliminate $\mathbf{a}_i$, define the $T \times T$ matrix $\mathbf{M} = \mathbf{I}_T - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$, and premultipy (2.6) by $\mathbf{M}$:

$$\mathbf{M}\mathbf{y}_i = (\mathbf{M}\mathbf{X}_i)\boldsymbol{\beta} + \mathbf{M}\mathbf{v}_i = (\mathbf{M}\mathbf{X}_i)\boldsymbol{\beta} + (\mathbf{M}\mathbf{X}_i)\mathbf{d}_i + \mathbf{M}\mathbf{u}_i.$$

We can write the equation in terms of residuals from individual-specific regressions as

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{v}}_i = \ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{X}}_i\mathbf{d}_i + \ddot{\mathbf{u}}_i \tag{2.7}$$

or

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{v}_{it}, t = 1, \ldots, T, \tag{2.8}$$

where, for instance, $\ddot{\mathbf{x}}_{it}$ is the $1 \times K$ vector of residuals from the regression $\mathbf{x}_{it}$ on $\mathbf{w}_t, t = 1, \ldots, T$. The fixed effects (FE) estimator of $\boldsymbol{\beta}$ – interpreted in the general sense of eliminating $\mathbf{a}_i$ from (2.1) – is just the pooled OLS estimator from (2.8). Rather than just restricting attention to time-demeaning, as in the usual fixed effects analysis, we allow for very general kinds of individual-specific "detrending."

Since the FE estimator, $\hat{\boldsymbol{\beta}}$, is just a pooled OLS estimator, sufficient conditions for consistency are simple to obtain. In addition to the rank condition

$$\text{rank } \mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i) = K, \tag{2.9}$$

a sufficient condition is

$$\mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{v}}_i) = \mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i) + \mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{u}}_i) = \mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i) + \mathrm{E}(\ddot{\mathbf{X}}_i'\mathbf{u}_i) = \mathbf{0}.$$

Now, by (2.2), $\mathrm{E}(\mathbf{u}_i|\ddot{\mathbf{X}}_i) = \mathbf{0}$, and so we must only worry about $\mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i)$. If

$$\mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i) = \mathbf{0} \tag{2.10}$$

then the FE estimator will be consistent. Since $\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i = \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}$, a sufficient condition is

$$\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\mathbf{d}_i) = \mathbf{0}, t = 1, \ldots, T. \tag{2.11}$$

Conditions (2.10) and (2.11) are a bit difficult to interpret. A simpler condition that is sufficient for (2.11) is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i), t = 1, \ldots, T, \tag{2.12}$$

which says that $\mathbf{b}_i$ is mean independent of all of the "detrended" $\mathbf{x}_{it}$. [If we slightly strengthen (2.12) to $\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{i1}, \ldots \ddot{\mathbf{x}}_{iT}) = \mathrm{E}(\mathbf{b}_i)$, then the fixed effects estimator can be shown to be unbiased, provided the expectation exists.] Condition (2.12) is notably weaker than the standard assumption assumed in a random effects environment, that $\mathbf{b}_i$ is mean independent of each $\mathbf{x}_{it}$. Intuitively, condition (2.12) allows $\mathbf{b}_i$ to be correlated with systematic components of $\mathbf{x}_{it}$. We give some specific examples in Section 3.

Generally, (2.12) is more likely to hold the richer is $\mathbf{w}_t$. So, even of we do not think the term $\mathbf{w}_t\mathbf{a}_i$ is necessary in (2.1), acting as if (2.1) contains individual-specific trends affords more robustness for estimating $\boldsymbol{\beta}$ because more individual specific features are swept out of $\mathbf{x}_{it}$. Of course, the more that is included in $\mathbf{w}_t$, the less variation there is in $\{\ddot{\mathbf{x}}_{it} : t = 1, \ldots, T\}$, and so efficiency of $\boldsymbol{\beta}$ can be adversely affected. In the limiting case $J = T$, $\ddot{\mathbf{x}}_{it} = \mathbf{0}, t = 1, \ldots, T$, and the fixed effects procedure cannot be carried out.

Estimating the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is straightforward with large N and small T. The usual, fully robust estimator – for example, Wooldridge [2002a, equation (10.59)] – is consistent:

$$\widehat{Avar(\hat{\boldsymbol{\beta}})} = \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\ddot{\mathbf{X}}_i\right)\left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}, \tag{2.13}$$

where $\hat{\mathbf{u}}_i \equiv \ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\hat{\boldsymbol{\beta}}$ are the $T \times 1$ vectors of fixed effects residuals. Even if we assume homoskedasticity and serial independence of $\{u_{it} : t = 1, \ldots, T\}$ [conditional on $(\mathbf{X}_i, \mathbf{a}_i, \mathbf{b}_i)$], a fully robust variance matrix is needed if $\mathbf{b}_i \neq \boldsymbol{\beta}$: the presence of $\ddot{\mathbf{X}}_i\mathbf{d}_i$ in the error terms induces both conditional heteroskedasticity and serial dependence. Fortunately, (2.13) is computed routinely by many regression packages, sometimes under the description of a "cluster-robust variance matrix estimator."

3. Methods Based on First Differencing

Often in empirical work, first differencing is used in place of the within transformation in order to eliminate an additive, unobserved effect. It is easy to see that the first difference (FD) estimator has robustness properties similar to the FE estimator.

In the model with a single additive, unobserved effect in $\mathbf{a}_i$, first differencing gives

$$\triangle y_{it} = \triangle\mathbf{x}_{it}\boldsymbol{\beta} + \triangle\mathbf{x}_{it}\mathbf{d}_i + \triangle u_{it}, t = 2, \ldots, T. \tag{3.1}$$

Using an argument similar to the fixed effects case, under a standard rank condition and (2.2), a sufficient condition for consistency of the FD estimator is

$$\mathrm{E}(\mathbf{b}_i|\triangle\mathbf{x}_{it}) = \mathrm{E}(\mathbf{b}_i), t = 2, \ldots, T, \tag{3.2}$$

which explicitly allows $\mathbf{b}_i$ to be correlated with the first-period covariates, $\mathbf{x}_{i1}$. When $T = 2$ and , (2.12) and (3.2) are the same condition when $\ddot{\mathbf{x}}_{it}$ are the time-demeaned covariates. When $T > 2$, (3.2) differs from (2.12), but they are similar in flavor. A fully robust asymptotic variance matrix estimator for the FD estimator can be routinely computed after pooled OLS

estimation in first differences. See Wooldridge (2002a, Section 10.6.2).

For more complicated models, first differencing can be followed by a fixed effects type analysis to eliminate additional unobserved heterogeneity, in which case the model in first differences can be analyzed as in Section 2. We explicitly cover the random trend model in the next section.

4. Some Examples

4.1. The Basic Additive Model

As mentioned earlier, a special case of the setup in Section 2 is the usual unobserved effects model estimated by fixed effects. Then, $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, where $\bar{\mathbf{x}}_i = \sum_{r=1}^{T} \mathbf{x}_{ir}$. Condition (2.12) means that $\mathbf{b}_i$ can be correlated with $\bar{\mathbf{x}}_i$ provided that $\mathbf{b}_i$ is conditionally mean independent of the deviations from the means, $\ddot{\mathbf{x}}_{it}$. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$, then (2.12) allows for arbitrary correlation between $\mathbf{f}_i$ and $\mathbf{b}_i$, provided

$$\mathrm{E}(\mathbf{b}_i | \mathbf{r}_{i1}, \ldots, \mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i) \tag{4.1}$$

Similarly, the first differencing estimator is also consistent under (4.1); it, too, allows arbitrary correlation between $\mathbf{b}_i$ and $\mathbf{f}_i$.

4.2. Random Trend Models

If we specify (2.1) as a random trend model, there are two popular approaches to estimation. The pure fixed effects approach is to follow the procedure from Section 2 – so that the $\ddot{\mathbf{x}}_{it}$ are the detrended values from the regression $\mathbf{x}_{it}$ on $1, t, t = 1, \ldots, T$, for each i. Then, we can allow even more dependence between $\mathbf{b}_i$ and time-constant features of $\mathbf{x}_{it}$. For example, suppose we can write

$$\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{g}_i t + \mathbf{r}_{it}, t = 1, \ldots, T, \tag{4.2}$$

so that each element of $\mathbf{x}_{it}$ is allowed to have an individual-specific trend. Then, for each i, $\ddot{\mathbf{x}}_{it}$ depends only on $\{\mathbf{r}_{i1}, \ldots, \mathbf{r}_{iT}\}$, and so (4.1) is again sufficient. In applications of (2.1), we are usually worried that $\mathbf{b}_i$ is correlated with time-constant components of $\mathbf{x}_{it} - \mathbf{f}_i$ and $\mathbf{g}_i$ in the case of (4.2) – in which case (4.1) seems reasonable. The process in (4.2) includes the case where $\mathbf{x}_{it}$ is an integrated of order one process with individual-specific drift, as in

$$\mathbf{x}_{it} = \mathbf{g}_i + \mathbf{x}_{i,t-1} + \mathbf{q}_{it}, t = 1, \ldots, T, \tag{4.3}$$

where $\{\mathbf{q}_{it} : t = 1, \ldots, T\}$ can have arbitrary serial correlation. Repeated substitution shows that (4.2) holds with $\mathbf{f}_i = \mathbf{x}_{i0}$ and $\mathbf{r}_{it} = \sum_{s=1}^{t} \mathbf{q}_{is}$. Since $\{\mathbf{r}_{it} : t = 1, \ldots, T\}$ is a function of $\{\mathbf{q}_{it} : t = 1, \ldots, T\}$, (4.1) holds if $\mathrm{E}(\mathbf{b}_i | \mathbf{q}_{i1}, \ldots, \mathbf{q}_{iT}) = \boldsymbol{\beta}$, which seems reasonable since we can allow $\mathbf{b}_i$ to be arbitrarily correlated with the vector of initial conditions, $\mathbf{x}_{i0}$, as well as the vector of drifts, $\mathbf{g}_i$.

An alternative estimation approach is to first difference to eliminate the additive effect, and then to use the within transformation to account for the random trend. First differencing is more attractive than the pure fixed effects approach from Section 2 when $\{u_{it} : t = 1, \ldots, T\}$ contains substantial positive serial correlation. Since we are applying the within

transformation to the first differenced equation, we see that a sufficient condition for consistency is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\triangle}\mathbf{x}_{it}) = \mathrm{E}(\mathbf{b}_i), t = 2, \ldots, T, \tag{4.4}$$

where $\ddot{\triangle}\mathbf{x}_{it}$ denotes the time-demeaned first differences. If $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ follows (4.2), then first differencing $\mathbf{x}_{it}$ eliminates $\mathbf{f}_i$ while the within transformation applied to the first differences eliminates $\mathbf{g}_i$. In other words, (4.1) is still sufficient for consistency.

Similar conclusions hold for both FE and strategies based on differencing if we take $\mathbf{w}_t = (1, t, t^2)$ (provided $T \geq 4$). Then, $\mathbf{x}_{it}$ can have an individual-specific quadratic trend, provided $\mathbf{b}_i$ is mean independent of the idiosyncratic part of $\mathbf{x}_{it}$. And so on.

4.3. Estimating Average Treatment Effects with Unobserved Heterogeneity

The results in Sections 2 and 3 have interesting implications for estimating average treatment effects (ATEs) in a class of nonlinear unobserved effects panel data models. For motivation, consider an example due to Hahn (2001), who was commenting on Angrist (2001). Hahn (2001) considered an unobserved effects probit model with two periods of panel data, and a single binary treatment indicator, x_{it}:

$$\mathrm{P}(y_{it} = 1|x_{i1}, x_{i2}, c_i) = \Phi(c_i + \gamma x_{it}), t = 1, 2, \tag{4.5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Hahn also assumed that y_{i1} and y_{i2} are independent conditional on (x_{i1}, x_{i2}, c_i), and that no units are treated in the first time period while all are treated in the second: $(x_{i1}, x_{i2}) = (0, 1)$. The last assumption implies

that (x_{i1}, x_{i2}) is independent of c_i, which would seem to be ideal for estimating the only parameter in the model, γ. Hahn points out that, even with all of the assumptions he imposes, γ is not known to be identified. On the other hand, the average treatment effect, $\beta \equiv \mathrm{E}[\Phi(c_i + \gamma) - \Phi(c_i)]$, is identified, and a simple, consistent estimator is $\hat{\beta} \equiv N^{-1} \sum_{i=1}^{N} (y_{i2} - y_{i1})$. It is easy to see that $\hat{\beta}$ is the usual fixed effects estimator in the simple linear model $y_{it} = a_i + \beta x_{it} + u_{it}, t = 1, 2$. (Recall that FE is identical to FD when $T = 2$, and the FD estimator is easily seen to be $\hat{\beta}$ because $x_{i2} - x_{i1} = 1$ for all i.). Hahn (2001) uses this example to show that ATEs can be identified even when underlying parameters are probably not. But he also uses the special structure of $\{x_{it} : t = 1, 2\}$ to argue that the success of Angrist's (2001) strategy of eschewing nonlinear models in favor of linear methods – even when y_{it} is a limited dependent variable – hinges on the structure of treatment assignment. Here, I use the results from Section 2 to determine assumptions under which simple panel data strategies do recover average treatment effects.

We can identify average treatment effects in a very general class of unobserved effects models, provided we make assumptions of the kind in Section 2, and assume no time heterogeneity. Consider

$$\mathrm{E}(y_{it} | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = h(\mathbf{x}_{it}, \mathbf{c}_i), t = 1, \ldots, T, \tag{4.6}$$

where $h(\cdot, \cdot)$ is an unknown function, $\mathbf{c}_i$ is a vector of unobserved heterogeneity, and $\mathbf{x}_{it}$ is a $1 \times K$ vector of mutually exclusive binary "treatment" indicators. This structure for $\mathbf{x}_{it}$ is very common in the treatment effect literature, where the base group (in time period t) is characterized by $\mathbf{x}_{it} = (x_{it1}, x_{it2}, \ldots, x_{itK}) = \mathbf{0}$. Other units in the population are subjected to one, and only one, of K treatments. For example, in the population of people with at least a

high school education, the base group could be people with no additional schooling. The treatment indicators can denote different amounts of college. Or, perhaps people participate in a job training program at different levels, with $\mathbf{x}_{it} = \mathbf{0}$ indicating no job training. The leading case is $K = 1$, where x_{it} is a binary treatment indicator.

There are only two assumptions in (4.6). The first is strict exogeneity of the treatment indicators, $\mathbf{x}_{it}$, conditional on $\mathbf{c}_i$. We have maintained strict exogeneity throughout, and it is very difficult to relax in general unobserved effects models. Second, (4.6) implies that the treatment effects are constant across time. For cross sectional unit i, the treatment effect of treatment level j (relative to no treatment) is

$$b_{ij} \equiv h(\mathbf{e}_j, \mathbf{c}_i) - h(\mathbf{0}, \mathbf{c}_i), \tag{4.7}$$

where $\mathbf{e}_j$ is the vector with one in its j^{th} entry and zeros elsewhere. Therefore, the ATEs are

$$\beta_j = \mathrm{E}[h(\mathbf{e}_j, \mathbf{c}_i) - h(\mathbf{0}, \mathbf{c}_i)] = \mathrm{E}(b_{ij}), j = 1, \ldots, K. \tag{4.8}$$

The goal is to determine when the usual fixed effects estimator, applied to a linear model, consistently estimates the ATEs. This is simple in the pure treatment effects setup because we can write

$$\mathrm{E}(y_{it}|\mathbf{X}_i, \mathbf{c}_i) = a_i + \mathbf{x}_{it}\mathbf{b}_i, t = 1, \ldots, T \tag{4.9}$$

where $a_i \equiv h(\mathbf{0}, \mathbf{c}_i)$ and $\mathbf{b}_i$ is the $K \times 1$ vector of individual-specific treatment effects, b_{ij}. Equation (4.9) holds because each cross-sectional unit falls into one, and only one, treatment class at time t. Given (4.9), we can apply the results for the fixed effects estimator from Section 2. If $\mathbf{c}_i$ is independent of the time-demeaned covariates, $\{\ddot{\mathbf{x}}_{it} : t = 1, \ldots, T\}$, then so is $\mathbf{b}_i$, and condition (2.12) holds. It follows that, regardless of the nature of y_{it}, for any pattern of serial dependence, and for general treatment patterns over time – even some that induce

correlation between $\mathbf{x}_{it}$ and $\mathbf{c}_i$ – the FE estimator consistently estimates the average treatment effects. Similar comments hold for the first differencing estimator.

Unfortunately, model (4.6) is not as general as we would like. For one, it does not allow other individual-specific covariates to affect y_{it}. Perhaps most importantly, (4.6) excludes aggregate time effects, which generally allow ATEs to vary with time, and can be important in policy analysis with panel data. It turns out that we can identify, and easily estimate, time-varying ATEs in a general model, provided we change the assumption about the relationship between the unobserved heterogeneity and $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$. For simplicity, let x_{it} be a binary treatment indicator, and replace (4.6) with

$$\mathrm{E}(y_{it}|x_{i1}, \ldots, x_{iT}, \mathbf{c}_i) = h_t(x_{it}, \mathbf{c}_i), t = 1, \ldots, T, \tag{4.10}$$

so that $h_t(\bullet, \bullet)$ is allowed to vary with time. The average treatment effect now depends on t:

$$\beta_t = \mathrm{E}[h_t(1, \mathbf{c}_i) - h_t(0, \mathbf{c}_i)], t = 1, \ldots, T, \tag{4.11}$$

Now, rather than assuming that $\mathbf{c}_i$ is independent of $\{\ddot{\mathbf{x}}_{it} : t = 1, \ldots, T\}$, we assume independence conditional on $\bar{x}_i$:

$$\mathrm{D}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT}) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i) \text{ or } \mathrm{D}(\mathbf{c}_i|\bar{x}_i, \ddot{x}_{i1}, \ldots, \ddot{x}_{iT}) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i). \tag{4.12}$$

Assumption (4.12) is a nonparametric version of Mundlak's (1978) conditional mean assumption in the linear case; see also Chamberlain (1984) and Wooldridge (2002a). It states that the distribution of the unobserved effect, given the observed history of treatments, depends only on the fraction of periods treated. Condition (4.12) is similar in spirit to (2.12), but it is not the same, even if (4.12) could be stated in terms of conditional expectations. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$ and $\mathbf{c}_i \equiv \mathbf{b}_i$, (4.1) is sufficient for (2.12), but (4.1) does not imply $\mathrm{E}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT}) = \mathrm{E}(\mathbf{c}_i|\bar{x}_i)$.

Under (4.10) and (4.12) we have

$$\begin{aligned} \mathrm{E}(y_{it}|\mathbf{X}_i) &= \int h_t(x_{it},\mathbf{c})dG(\mathbf{c}|\mathbf{X}_i) = \int h_t(x_{it},\mathbf{c})dG(\mathbf{c}|\bar{x}_i) \equiv m_t(x_{it},\bar{x}_i) \\ &= \mathrm{E}(y_{it}|x_{it},\bar{x}_i), t = 1,\ldots,T. \end{aligned} \tag{4.13}$$

The key is that $\mathrm{E}(y_{it}|\mathbf{X}_i)$ does not depend on $\{x_{i1},\ldots,x_{iT}\}$ in an unrestricted fashion. If x_{it} were continuous, or took on numerous values, we could use nonparametric methods to estimate $m_t(\cdot,\cdot)$. In the treatment effect case, estimation is very simple because $(x_{it},\bar{x}_i)$ can take on only $2(T+1)$ different values (since x_{it} takes on only two values and $\bar{x}_i$. takes on the values $\{0,1/T,\ldots,(T-1)/T,1\}$). Let $s_{i1} = 1[\bar{x}_i = 1/T]$, $s_{i2} = 1[\bar{x}_i = 2/T]$, …, and $s_{iT} = 1[\bar{x}_i = 1]$. Then we can write

$$\mathrm{E}(y_{it}|\mathbf{X}_i) = \alpha_t + \beta_t x_{it} + \mathbf{s}_i\boldsymbol{\gamma}_t + x_{it}(\mathbf{s}_i - \boldsymbol{\mu}_s)\boldsymbol{\delta}_t, t = 1,\ldots,T \tag{4.14}$$

where $\mathbf{s}_i$ is the $1 \times T$ vector of s_{it} and $\boldsymbol{\mu}_s \equiv \mathrm{E}(\mathbf{s}_i)$. The coefficient on x_{it} is the average treatment effect. [Generally, iterated expectations implies that $\beta_t = \mathrm{E}[\mathrm{E}[h_t(1,\mathbf{c}_i) - h_t(0,\mathbf{c}_i)|\bar{x}_i]] = \mathrm{E}[m_t(1,\bar{x}_i) - m_t(0,\bar{x}_i)]$; see Wooldridge (2002b, Lemma 2.2) for a general treatment.] Subtracting $\boldsymbol{\mu}_s$ from $\mathbf{s}_i$ before forming the interactions ensures β_t is the treatment effect. In practice, $\boldsymbol{\mu}_s$ would be replaced with $\bar{\mathbf{s}} = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i$. In other words, for each period t, we run the regression

$$y_{it} \text{ on } 1, x_{it}, s_{i1},\ldots,s_{iT}, x_{it}(s_{i1} - \bar{s}_1),\ldots,x_{it}(s_{iT} - \bar{s}_T), i = 1,\ldots,N, \tag{4.15}$$

where the coefficient $\hat{\beta}_t$ on x_{it} is the estimated ATE for period t.

If we made the random effects assumption $\mathrm{D}(\mathbf{c}_i|\mathbf{X}_i) = \mathrm{D}(\mathbf{c}_i)$ then, of course, the simple regression of y_{it} on $1, x_{it}, i = 1,\ldots N$ would consistently estimate β_t. If we pool across t (as well as i) and run the regression y_{it} on $1, d2_t,\ldots,dT_t, x_{it}, \bar{x}_i, t = 1,\ldots,T; i = 1,\ldots,N$, where dr_t is a period r dummy variable, then the common coefficient on x_{it}, which is identical to the

fixed effects estimate, would be the estimate of the ATE (assumed constant across t). The regression in (4.15) is more flexible because it allows ATEs to change over time while allowing $\mathrm{E}(y_{it}|x_{it},\bar{x}_i)$ to depend on $(x_{it},\bar{x}_i)$ in a completely general way. Provided $\{x_{it} : t = 1,\ldots,T\}$ has some time variation, x_{it} and $\bar{x}_i$ will have independent variation for any t, which is all we need to identify β_t under (4.12).

Condition (4.12) is hardly general, but it can be relaxed with $T > 2$. For example, if $\overline{\Delta x}_i \equiv (T-1)^{-1}\sum_{t=2}^{T}\Delta x_{it}$ is the average change in treatment over the T periods, we might replace (4.12) with

$$\mathrm{D}(\mathbf{c}_i|\mathbf{X}_i) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i,\overline{\Delta x}_i) \text{ or } \mathrm{D}(\mathbf{c}_i|x_{i1},\ldots,x_{iT}) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i,x_{iT}-x_{i1}), \tag{4.16}$$

where equivalence follows because $\overline{\Delta x}_i = (x_{iT}-x_{i1})/(T-1)$. [This assumption is in the spirit of assuming $\mathbf{b}_i$, the vector of slopes in a linear model, is independent of $\{\mathbf{r}_{i1},\ldots,\mathbf{r}_{iT}\}$ in (4.2); but it is not the same condition.] Then,

$$\mathrm{E}(y_{it}|\mathbf{X}_i) = \mathrm{E}(y_{it}|x_{it},\bar{x}_i,\overline{\Delta x}_i) \equiv m_t(x_{it},\bar{x}_i,\overline{\Delta x}_i). \tag{4.17}$$

Except for special treatment patterns, the ATE for each time period is identified from the population regression of y_{it} on $(x_{it},\bar{x}_i,\overline{\Delta x}_i)$ provided $T \geq 3$. Generally, the regressors in each time period can take on $2 \cdot (T+1) \cdot 3 = 6(T+1)$ different values because $\overline{\Delta x}_i$ takes on values in $\{-1/(T-1),0,1/(T-1)\}$. We can estimate a saturated regression model by defining two dummy variables, say, w_{i0},w_{i1}, for $\overline{\Delta x}_i$ taking on the values 0 and $1/(T-1)$, respectively. For each time period t, the regression would contain an overall intercept, $x_{it},\mathbf{s}_i,\mathbf{w}_i$, interactions $s_{ij}w_{ik}$, and interactions $x_{it}(s_{ij}-\bar{s}_j)$,$x_{it}(w_{ik}-\bar{w}_k)$, and $x_{it}(s_{ij}w_{ik}-\overline{s_jw_k})$ for all j and k. Demeaning all of the indicators, including $s_{ij}w_{ik}$, before forming the interactions with x_{it}, ensures that the coefficient on x_{it} is the average treatment effect. As in many cases, it makes

sense to obtain heteroskedasticity-robust standard errors for the ATEs.

The procedure described in the previous paragraph is costly in terms of degrees of freedom. For one, T different cross-sectional regressions are used; there is not pooling across t. So, for estimating β_t, one has $N - 6(T + 1)$ degrees of freedom. The panel structure of the data is used only in obtaining the time-constant controls, s_{ij} and w_{ik}. One could use as regressors $1, x_{it}, \bar{x}_i, x_{it} - x_{i1}, x_{it}(\bar{x}_i - \hat{\mu}_{\bar{x}}), x_{it}(\overline{\Delta x}_i - \hat{\mu}_{\overline{\Delta x}})$, where $\hat{\mu}_{\bar{x}}$ is the cross-sectional average of $\bar{x}_i$ and $\hat{\mu}_{\overline{\Delta x}}$ is the average of $\overline{\Delta x}_i$. Still, there is something to say for the general procedure, as it may properly reflect the uncertainty in estimating ATEs under nonparametric assumptions.

While further embellishments are possible with large T, identification of the ATEs in every time period hinges on the functions of $\{x_{it} : t = 1, \ldots, T\}$ assumed to appear in $\mathrm{D}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT})$. We cannot allow $\mathrm{D}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT})$ to be entirely unrestricted.

How do the above procedures compare with more common approaches? A general comparison is not possible because (4.10) puts very little structure on $\mathrm{E}(y_{it}|\mathbf{X}_i, \mathbf{c}_i)$ [at the cost of (4.12) or (4.16)]. But suppose y_{it} is a binary response:

$$\mathrm{P}(y_{it} = 1|\mathbf{X}_i, c_i) = F(\delta_t + \gamma x_{it} + c_i), t = 1, \ldots, T, \tag{4.18}$$

where $F(\cdot)$ is a cumulative distribution function. If we take F to be the logistic function, and the y_{it} are conditionally independent across time, then the fixed effects logit estimator is consistent for γ (and the aggregate time effect coefficients). Unfortunately, ATEs are not identified since we make no distributional assumption for c_i. Essentially by construction, methods that take no stand concerning the unconditional distribution of c_i, or the conditional distribution $D(c_i|\mathbf{X}_i)$, have little hope of identifying ATES.

If F is the standard normal cdf, Chamberlain's (1980) random effects probit model can be used, provided we assume $c_i|\mathbf{X}_i \sim \mathrm{Normal}(\xi_0 + \boldsymbol{\xi}_1 x_{i1} + \ldots + \boldsymbol{\xi}_T x_{iT}, \eta^2)$. [In principle, F could be

the logit function, but then implementation of Chamberlain's method is much more difficult.] Chamberlain's approach identifies γ as well as the ATEs – see Chamberlain (1984) or Wooldridge (2002a, Chapter 15) – the latter of which vary over time because of the presence of δ_t. Compared with the procedure discussed above, Chamberlain's method allows unrestricted weights on the x_{it} in $\mathrm{E}(c_i|\mathbf{X}_i)$, at the cost of homoskedasticity and normality. The regression procedure outlined above replaces Chamberlain's parametric assumptions with (4.12) or (4.16). The two approaches are complementary, since they work under different sets of assumptions, neither of which nests the other.

All of the methods described above can be extended to the case of $K + 1$ treatment levels, but degrees of freedom could be an issue. Then, each of the K elements in $\bar{\mathbf{x}}_i$ can take on $T + 1$ different values, and so $K(T + 1)$ dummy variables are needed to saturate the model, and these each need to be interacted with the elements of $\mathbf{x}_{it}$. A large cross-sectional sample would be needed to implement a fully nonparametric analysis under (4.12), and the extension in (4.16) would require even more data.

5. Other Extensions

Sometimes, we want to allow $\mathbf{b}_i$ to vary with observed, time-constant covariates, say $\mathbf{z}_i$, a $1 \times L$ vector:

$$\mathbf{b}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{z}_i' + \mathbf{d}_i, \tag{5.1}$$

where $\boldsymbol{\alpha}$ is $K \times 1$ and $\boldsymbol{\Gamma}$ is $K \times L$. (One possibility is to include the time averages, $\bar{\mathbf{x}}_i$, in $\mathbf{z}_i$.) Under (4.1), we can write

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\boldsymbol{\beta} + (\mathbf{z}_i \otimes \mathbf{x}_{it})\boldsymbol{\gamma} + v_{it}, t = 1, \ldots, T, \tag{5.2}$$

where $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma})$ and $v_{it} = \mathbf{x}_{it}\mathbf{d}_i + u_{it}$, as before. Equation (5.2) is just the formal way of writing that we add to the original model interactions between the elements of $\mathbf{z}_i$ and $\mathbf{x}_{it}$. If $\text{E}(\mathbf{d}_i|\ddot{\mathbf{x}}_{it}) = \text{E}(\mathbf{d}_i) = \mathbf{0}$, the fixed effects estimator applied to (5.2) would consistently estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, in which case the average effects $\boldsymbol{\beta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\text{E}(\mathbf{z}_i')$ are consistently estimated by $\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\Gamma}}\bar{\mathbf{z}}'$, where $\bar{\mathbf{z}}$ is the sample average across i.

The methods from Section 4.3 can also be extended when time constant covariates, $\mathbf{z}_i$, are available. For example, in (4.10), we could replace (4.12) with

$$\text{D}(\mathbf{c}_i|\mathbf{X}_i, \mathbf{z}_i) = \text{D}(\mathbf{c}_i|\bar{x}_i, \mathbf{z}_i), \tag{5.3}$$

in which case (4.13) becomes $\text{E}(y_{it}|\mathbf{X}_i, \mathbf{z}_i) = \text{E}(y_{it}|x_{it}, \bar{x}_i, \mathbf{z}_i), t = 1, \ldots, T$. An estimate of the ATE at time t is obtained as $N^{-1}\sum_{i=1}^{N}[\hat{\text{E}}(y_{it}|1, \bar{x}_i, \mathbf{z}_i) - \hat{\text{E}}(y_{it}|0, \bar{x}_i, \mathbf{z}_i)]$, for a suitable estimator of the conditional expectation. By including in $\mathbf{z}_i$ observables such as family background, education, pre-training earnings, and so on, the assumption that heterogeneity depends only on the average treatment level may be more plausible. A thorough study that considers estimating $\text{E}(y_{it}|x_{it}, \bar{x}_i, \mathbf{z}_i)$ when the dimension of $\mathbf{z}_i$ is large or contains continuous variables is left for future study.

References

Angrist, J.D. (2001), " Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and*

Economic Statistics 19, 2-16.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225-238.

Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics*, Volume 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North Holland, 1247-1318.

Friedberg, L. (1998), "Did Unilateral Divorce Raise Divorce Rates?" *American Economic Review* 88, 608-627.

Hahn, J. (2001), "Comment: Binary Regressors in Nonlinear Panel-Data Models with Fixed Effects," *Journal of Business and Economic Statistics* 19, 16-17.

Heckman, J.J. and V.J. Hotz (1989), "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association* 84, 862-874.

Heckman, J.J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model," *Journal of Human Resources* 33, 974-987.

Hoxby, C.M. (1996), "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics* 111, 671-718.

Hsiao, C. (1986), *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Mundlak, Y. (1978), On the Pooling of Time Series and Cross Section Data," *Econometrica* 46, 69-85.

Papke, L.E. (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37-49.

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

Wooldridge, J.M. (2002b), "Unobserved Heterogeneity and Estimation of Average Partial Effects," mimeo, Michigan State University Department of Economics.

Wooldridge, J.M. (2003), “Fixed Effects Estimation of the Population-Averaged Slopes in a Panel Data Random Coefficient Model,” forthcoming, Problem, *Econometric Theory* 19.