FIXED-EFFECTS AND RELATED ESTIMATORS FOR CORRELATED RANDOM-COEFFICIENT AND TREATMENT-EFFECT PANEL DATA MODELS

Jeffrey M. Wooldridge*

Abstract—I derive conditions under which a class of fixed-effects estimators consistently estimates 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. I apply the results, and propose alternative estimators, to estimation of average treatment in a class of nonlinear unobserved-effects models.

I. 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 freely correlated with the time-varying covariates. (On the other hand, with random-effects methods we assume that unobserved heterogeneity is uncorrelated with observed covariates.) An important extension of the standard linear model with an additive unobserved effect is the random trend model, where each cross-sectional unit is allowed to have its own linear trend (in addition to a separate level effect); see, for example, Heckman and Hotz (1989). Wooldridge (2002, section 11.2) provides an overview of these kinds of models.

The properties of fixed-effects estimators have been derived assuming constant coefficients on the individual-specific, time-varying covariates. Recently, there has been an upsurge of interest in random-coefficient models, where the slope coefficients can vary with the cross-sectional unit. Part of this interest stems from models with heterogeneous treatment effects. [The literature is too vast to review here; for a list of references, see Wooldridge (2002, chapter 18).] In such models, interest typically centers on the population-averaged effect, or, in the specific context of causal effects, the average treatment effect. An interesting question is: If we use methods that assume constant partial effects, do we nevertheless estimate interesting population parameters if the true relationship has individual-specific slopes? In Wooldridge (2003), I pointed out that the usual fixed-effects estimator in the standard additive model is consistent for the population-averaged effect in a model with individual-specific slopes whenever the slopes are mean-independent of the time–de-meaned covariates. As we show in section 4, this finding means that the individual-specific slopes can be correlated with time-constant features of the covariates.

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, which seems unrealistic for most economic applications. Hsiao (1986, chapter 6) gives a detailed treatment.

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 individual-specific trends is satisfyingly robust to the presence of individual-specific slopes on the individual-specific covariates. Essentially, the slopes can be correlated with level and trend features of the covariates; I give precise conditions in sections III and IV.

The remainder of the paper is organized as follows. Section II contains a motivating example due to Hahn (2001). Section III contains the main result, and section IV shows how the results apply to the usual additive unobserved-effects model, as well as the random-trend model, which has become popular in empirical studies [for example, Papke (1994) and Friedberg (1998)]. I extend Hahn's (2001) model to allow for a general index function, serial dependence, and general treatment patterns, and I propose modified estimators that consistently estimate time-varying average treatment effects.

II. An Example

To motivate our interest in fixed-effects-type estimators in models with invididual-specific slopes, I start with an example due to Hahn (2001), who was commenting on Angrist (2001). Hahn considered an unobserved-effects probit model with two periods of panel data and a single binary treatment indicator x_{it}:

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

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and c_i is unobserved heterogeneity. As pointed out by Hahn, the slope coefficient in the panel probit model is not known to be identified, unless we make assumptions about the distribution of c_i given (x_{i1},x_{i2}), denoted by $\mathrm{D}(c_i|x_{i1},x_{i2})$. Nevertheless, Hahn shows that it is possible to identify the average treatment effect (ATE), $\beta \equiv \mathrm{E}[\Phi(c_i + \gamma) - \Phi(c_i)]$, without specifying $\mathrm{D}(c_i|x_{i1},x_{i2})$. In particular, Hahn assumes 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

Received for publication July 8, 2003. Revision accepted for publication May 5, 2004.

* Michigan State University.

I would like to thank Jin Hahn, an anonymous referee, and Daron Acemoglu for helpful comments on an earlier draft. Thanks also to Josh Angrist for reminding me about the example in Jin Hahn's JBES comment. An earlier version of this paper circulated under the title "On the Robustness of Fixed Effects and Related Estimators in Correlated Random Coefficient Models."

The Review of Economics and Statistics, May 2005, 87(2): 385–390

whereas all are treated in the second: $(x_{i1},x_{i2}) = (0, 1)$. The latter assumption implies that (x_{i1},x_{i2}) is independent of c_i, but the distribution of c_i is unspecified. Under these assumptions, Hahn notes that $\hat{\beta} \equiv N^{-1}\sum_{i=1}^{N}(y_{i2} - y_{i1})$ is a consistent estimator of β. Interestingly, $\hat{\beta}$ is the usual fixed-effects (FE) estimator in the linear model $y_{it} = a_i + \beta x_{it} + e_{it}$, $t = 1,2$. The argument is simple: FE is identical to first-differencing (FD) when $T = 2$, and the FD estimator (ordinary least squares on the first differences) is easily seen to be $\hat{\beta}$, because $x_{i2} - x_{i1} = 1$ for all i.

For the purposes of the current paper, it is useful to write Hahn's model as a linear model with individual-specific intercepts and slopes:

$$y_{it} = a_i + b_i x_{it} + u_{it}, \qquad t = 1,2, \tag{2}$$

where $a_i \equiv \Phi(c_i)$, $b_i \equiv \Phi(c_i + \gamma) - \Phi(c_i)$, and $u_{it} \equiv y_{it} - \mathrm{E}(y_{it}|x_{i1},x_{i2},c_i)$, $t = 1,2$. Equation (2) is the simplest version of the model I treat in this paper. The previous discussion demonstrates that, in this simple example, the usual fixed-effects estimator that ignores the nonlinearity in (1)—and, therefore, the heterogeneous treatment effects—nevertheless consistently estimates the ATE, $\beta = \mathrm{E}(b_i)$.

Hahn used his example to show that ATEs are easily shown to be identified even when identification of underlying parameters is not obvious. But he also used the special setup as a potential caution: linear estimation methods such as fixed effects may have limited scope because the assumptions under which they are consistent for average effects are restrictive. In section IV I return to this example and show that Hahn's finding concerning consistency of the FE estimator is much more general.

III. A General Consistency Result for Fixed Effects in the Correlated Random-Coefficient Model

We now turn to analyzing a general random-coefficient panel data model. For a random draw i from the population, the model is

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it}, \qquad t = 1, \ldots, T, \tag{3}$$

where w_t is a $1 \times J$ vector of aggregate time variables—which we treat as nonrandom (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, $\mathbf{b}_i$ 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 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. In models where partial effects depend on unobservable heterogeneity, attention usually turns to the population-averaged effects, also called "average partial effects." In the example in section II, the average partial effect is simply the average treatment effect. A detailed discussion is given in Wooldridge (2005).

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

$$\mathrm{E}(u_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT},\mathbf{a}_i,\mathbf{b}_i) = 0, \qquad t = 1, \ldots, T, \tag{4}$$

which follows under the conditional-mean assumption

$$\begin{aligned}\mathrm{E}(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT},\mathbf{a}_i,\mathbf{b}_i) &= \mathrm{E}(y_{it}|\mathbf{x}_{it},\mathbf{a}_i,\mathbf{b}_i)\\ &= \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i, \qquad t = 1, \ldots, T.\end{aligned} \tag{5}$$

The assumption (5) is a standard strict exogeneity condition in unobserved-effects models: conditional on $(\mathbf{x}_{it},\mathbf{a}_i,\mathbf{b}_i)$, the covariates from the other time periods do not affect the expected value of y_{it}. Though this rules out the possibility of lagged dependent variables, it does not restrict the correlation between $(\mathbf{a}_i,\mathbf{b}_i)$ and $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$. It is important to explicitly condition on $(\mathbf{a}_i,\mathbf{b}_i)$ in equations (4) and (5) to emphasize that these represent unobserved heterogeneity that has some distribution in the population. The possibility of correlation between $\mathbf{b}_i$ and the $\mathbf{x}_{it}$ makes equation (3) 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 $\mathbf{b}_i = \boldsymbol{\beta}$ 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, for we then lose identification of $\boldsymbol{\beta}$. Generally, we must have $J < T$; see, for example, Wooldridge (2002, section 11.2). Aggregate time effects with constant coefficients—a standard feature of panel data models—are allowed. To keep the notation simple, we do not explicitly introduce time-period dummies.

One possibility for analyzing equation (3) is to treat the $\mathbf{a}_i$ and $\mathbf{b}_i$ as parameters to estimate for each i. Under equation (4) 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. Nevertheless, when estimation is possible, 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 (2002, section 11.2) for verification in a closely related context.] Because this strategy is not available for large K, and because 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 the model (3). Write $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$, where $\mathrm{E}(\mathbf{d}_i) = 0$ by definition. Simple substitution into equation (3) gives

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\boldsymbol{\beta} + (\mathbf{x}_{it}\mathbf{d}_i + u_{it}) \tag{6}$$

$$\equiv \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \tag{7}$$

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 equation (3), 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 tth row $\mathbf{w}_t$, let $\mathbf{X}_i$ be the $T \times K$ matrix with tth row $\mathbf{x}_{it}$, and let $\mathbf{v}_i$ be the vector of v_{it}. Then we can write

$$\mathbf{y}_i = \mathbf{W}\mathbf{a}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i = \mathbf{W}\mathbf{a}_i + \mathbf{X}_i\boldsymbol{\beta} + (\mathbf{X}_i\mathbf{d}_i + \mathbf{u}_i). \tag{8}$$

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 premultiply equation (8) by 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

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{v}}_i = \ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{X}}_i\mathbf{d}_i + \ddot{\mathbf{u}}_i, \tag{9}$$

or

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{v}_{it}, \qquad t = 1, \ldots, T, \tag{10}$$

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 FE estimator of $\boldsymbol{\beta}$—interpreted in the general sense of eliminating $\mathbf{a}_i$ from equation (3)—is just the pooled OLS estimator from equation (10). Rather than just restricting attention to time de-meaning as in the usual fixed-effects analysis, we allow for very general kinds of individual-specific detrending.

Because 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

$$\text{rank } \mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i) = K, \tag{11}$$

a sufficient condition is

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

Now, by equation (4), $\mathrm{E}(\mathbf{u}_i|\ddot{\mathbf{X}}_i) = 0$, and so we need only worry about $\mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i)$. If

$$\mathrm{E}(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\mathbf{d}_i) = 0, \tag{12}$$

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

$$\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\mathbf{d}_i) = 0, \qquad t = 1, \ldots, T. \tag{13}$$

The conditions (12) and (13) are a bit difficult to interpret. A simpler condition that is sufficient for equation (13) is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i), \qquad t = 1, \ldots, T, \tag{14}$$

which says that $\mathbf{b}_i$ is mean-independent of all of the detrended $\mathbf{x}_{it}$. [If we slightly strengthen equation (14) to $\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{i1}, \cdots \ddot{\mathbf{x}}_{iT}) = \mathrm{E}(\mathbf{b}_i)$, then the fixed-effects estimator can be shown to be unbiased, provided the expectation exists.] The condition (14) is notably weaker than the standard assumption in a random-effects environment, that $\mathbf{b}_i$ is mean-independent of each $\mathbf{x}_{it}$. Intuitively, the condition (14) allows $\mathbf{b}_i$ to be correlated with systematic components of $\mathbf{x}_{it}$; I give specific examples in section IV.

Because we are acting as if the coefficients on $\mathbf{x}_{it}$ were constant, we can include time period dummy variables in $\mathbf{x}_{it}$ to allow secular changes in both y_{it} and $\mathbf{x}_{it}$, an important feature of many empirical studies that apply fixed effects. The number of time dummies we include depends on the makeup of $\mathbf{w}_t$. In the usual fixed-effects case, where $w_t \equiv 1$, we can include $T - 1$ time dummies. If $\mathbf{w}_t$ also contains a linear trend, another time dummy gets dropped, because the model allows for individual-specific linear trends (and, therefore, an unrestricted aggregate trend). The condition (14) is unaffected, because the time period dummies are nonstochastic. The important thing to remember is that we are applying a standard fixed-effects procedure as if $\mathbf{b}_i = \boldsymbol{\beta}$, and so we can allow aggregate time effects in the usual way.

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

IV. Applications

A. The Basic Additive Model

As mentioned earlier, a special case of the setup in section III is the usual unobserved-effects model estimated by fixed effects. Then, $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, where $\bar{\mathbf{x}}_i = T^{-1}\sum_{r=1}^{T}\mathbf{x}_{ir}$. The condition (14) 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 equation (14) allows for arbitrary correlation between $\mathbf{f}_i$ and $\mathbf{b}_i$, provided

$$\mathrm{E}(\mathbf{b}_i|\mathbf{r}_{i1}, \ldots, \mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i). \tag{15}$$

The condition (15) means that $\mathbf{b}_i$ is mean-independent of the idiosyncratic deviations in $\mathbf{x}_{it}$. Importantly, equation (15) does not restrict the serial dependence in $\{\mathbf{r}_{i1}, \ldots, \mathbf{r}_{iT}\}$. For example, in a model to explain student performance (y_{it}) in

terms of spending (x_{it}), the school-specific effect of spending (b_i) can be correlated with historical attributes of schools (f_i) that lead to higher or lower spending, on average. The restriction is that the slope is not correlated with idiosyncratic deviations in spending.

B. Random-Trend Models

If we specify equation (3) as a random-trend model, so that $\mathbf{w}_t = (1,t)$ and $\mathbf{a}_i = (a_{i1},a_{i2})'$, then the fixed-effects estimator sweeps away the individual-specific trends, as well as the level effects. In other words, 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. Consequently, we can allow even more dependence between $\mathbf{b}_i$ and time-constant features of $\mathbf{x}_{it}$. For example, suppose we can write

$$\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{g}_i t + \mathbf{r}_{it}, \qquad t = 1, \ldots, T, \tag{16}$$

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 equation (15) is again sufficient.

In applications of equation (3), 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 equation (16)—in which case equation (15) seems reasonable. The process in equation (16) includes the case where $\mathbf{x}_{it}$ is an integrated order-1 process with individual-specific drift, as in

$$\mathbf{x}_{it} = \mathbf{g}_i + \mathbf{x}_{i,t-1} + \mathbf{q}_{it}, \qquad t = 1, \ldots, T, \tag{17}$$

where $\{\mathbf{q}_{it} : t = 1, \ldots, T\}$ can have arbitrary serial correlation. Repeated substitution shows that equation (16) holds with $\mathbf{f}_i = \mathbf{x}_{i0}$ and $\mathbf{r}_{it} = \sum_{s=1}^{t} \mathbf{q}_{is}$. Because $\{\mathbf{r}_{it} : t = 1, \ldots, T\}$ is a function of $\{\mathbf{q}_{it} : t = 1, \ldots, T\}$, equation (15) holds if $\mathrm{E}(\mathbf{b}_i|\mathbf{q}_{i1}, \ldots, \mathbf{q}_{iT}) = \beta$, which seems reasonable in that 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, $\mathbf{g}_i$. In the performance-spending example in section IV A, we can allow the school-specific slope that measures the effect of spending on performance, b_i, to be correlated with initial spending x_{i0} and a school-specific growth in spending, g_i.

Another popular approach to estimating the random-trend model is to first-difference to eliminate the additive effect a_{i1}, and then use the within transformation to eliminate the random trend a_{i2}. First-differencing is more attractive than the pure fixed-effects approach from section III when $\{u_{it} : t = 1, \ldots, T\}$ contains substantial positive serial correlation. Because we are applying the within transformation to the first-differenced equation, it is easy to see that sufficient condition for consistency is

$$\mathrm{E}(\mathbf{b}_i|\Delta\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i), \qquad t = 2, \ldots, T, \tag{18}$$

where $\Delta\ddot{\mathbf{x}}_{it} = \Delta\mathbf{x}_{it} - (T-1)^{-1}\sum_{r=2}^{T} \Delta\mathbf{x}_{ir}$ denotes the time–de-meaned first differences. If $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ follows equation (16), then first-differencing $\mathbf{x}_{it}$ eliminates $\mathbf{f}_i$, and the within transformation applied to the first differences eliminates $\mathbf{g}_i$. In other words, equation (15) 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. And so on.

C. Estimating Average Treatment Effects with Unobserved Heterogeneity

We now turn to a more general version of Hahn's (2001) example. It turns out that the fixed-effects estimator applied to a linear model identifies average treatment effects in a general class of nonlinear unobserved-effects models, provided we make assumptions of the kind in section III, and assume no time heterogeneity. Consider

$$\mathrm{E}(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT},\mathbf{c}_i) = h(\mathbf{x}_{it},\mathbf{c}_i), \qquad t = 1, \ldots, T, \tag{19}$$

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 "untreated" 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 $x_{itj} = 0$, $j = 1, \ldots, K$, indicating no job training, $x_{it1} = 1$ indicating the lowest level of training, and $x_{itK} = 1$ indicating the highest level of training. The leading case is $K = 1$, where x_{it} is a binary treatment indicator.

There are only two assumptions in equation (19). 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, equation (19) implies that the treatment effects are constant across time—an assumption we relax below. For cross-sectional unit i, the treatment effect of treatment level j (relative to no treatment) is

$$b_{ij} \equiv h(\mathbf{e}_j,\mathbf{c}_i) - h(\mathbf{0},\mathbf{c}_i), \tag{20}$$

where $\mathbf{e}_j$ is the vector with 1 in its jth entry and 0's elsewhere. Therefore, the ATEs are

$$\beta_j = \mathrm{E}[h(\mathbf{e}_j,\mathbf{c}_i) - h(\mathbf{0},\mathbf{c}_i)] = \mathrm{E}(b_{ij}), \quad j = 1, \ldots, K. \tag{21}$$

The goal is to determine when the usual fixed-effects estimator, applied to a linear model, consistently estimates the ATEs. But, as with the simple version in section II, we can write

$$\mathrm{E}(y_{it}|\mathbf{X}_i,\mathbf{c}_i) = a_i + \mathbf{x}_{it}\mathbf{b}_i, \qquad t = 1, \ldots, T, \tag{22}$$

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 (22) holds because each cross-sectional unit falls into one, and only one, treatment class at time t. Given equation (22), we can apply the results for the fixed-effects estimator from section III. If $\mathbf{c}_i$ is independent of the time–de-meaned covariates $\{\ddot{\mathbf{x}}_{it} : t = 1, \ldots, T\}$, then so is $\mathbf{b}_i$, and the condition (14) 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, the model (19) is not as general as we would like. Perhaps most importantly, it 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 equation (19) with

$$\mathrm{E}(y_{it}|x_{i1}, \ldots, x_{iT},\mathbf{c}_i) = h_t(x_{it},\mathbf{c}_i), \qquad t = 1, \ldots, T, \tag{23}$$

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

$$\beta_t = \mathrm{E}[h_t(1,\mathbf{c}_i) - h_t(0,\mathbf{c}_i)], \qquad t = 1, \ldots, T. \tag{24}$$

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

$$\begin{aligned} &\mathrm{D}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT}) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i) \quad \text{or} \\ &\quad \mathrm{D}(\mathbf{c}_i|\bar{x}_i,\ddot{x}_{i1}, \ldots, \ddot{x}_{iT}) = \mathrm{D}(\mathbf{c}_i|\bar{x}_i). \end{aligned} \tag{25}$$

This assumption is a nonparametric version of Mundlak's (1978) conditional-mean assumption in the linear case; see also Chamberlain (1984) and Wooldridge (2002, chapter 15). It states that the distribution of the unobserved effect, given the observed history of treatments, depends only on the fraction of periods treated. The condition (25) is similar in spirit to (14), but it is not the same, and would not be even if equation (25) could be stated in terms of conditional expectations. For example, if $x_{it} = f_i + r_{it}$, $t = 1, \ldots, T$, and $\mathbf{c}_i \equiv \mathbf{b}_i$, then equation (15) is sufficient for (14), but does not imply $\mathrm{E}(\mathbf{c}_i|x_{i1}, \ldots, x_{iT}) = \mathrm{E}(\mathbf{c}_i|\bar{x}_i)$—and therefore does not imply equation (25).

The condition (25) can be interpreted in terms of sufficient statistics. For a single cross-sectional draw i, we can think of the observations $\{x_{i1}, \ldots, x_{iT}\}$ as having a distribution depending on the "parameter vector" $\mathbf{c}_i$. Then equation (25) holds if the time average $\bar{x}_i$ is a sufficient statistic for $\mathbf{c}_i$. To see why, recall a simple relationship between conditional densities:

$$\begin{aligned} f(\mathbf{c}|x_1, \ldots, x_T,\bar{x}) &= f(\mathbf{c},x_1, \ldots, x_T|\bar{x})/f(x_1, \ldots, x_T|\bar{x}) \\ &= f(x_1, \ldots, x_T|\mathbf{c},\bar{x}) \cdot f(\mathbf{c}|\bar{x})/f(x_1, \ldots, x_T|\bar{x}), \end{aligned} \tag{26}$$

where the notation should be clear. But sufficiency of $\bar{x}_i$ for $\mathbf{c}_i$ means that $f(x_1, \ldots, x_T|\mathbf{c},\bar{x}) = f(x_1, \ldots, x_T|\bar{x})$, and plugging this into equation (26) gives $f(\mathbf{c}|x_1, \ldots, x_T,\bar{x}) = f(\mathbf{c}|\bar{x})$, which is equation (25). Because the x_{it} are binary responses, there is a well-known case where sufficiency holds: conditional on $\mathbf{c}_i$, $\{x_{it} : t = 1, \ldots, T\}$ is an independent, identically distributed Bernoulli sequence with response probability $p(\mathbf{c}_i)$—an unspecified function of $\mathbf{c}_i$.

Under equations (23) and (25) we have

$$\begin{aligned} \mathrm{E}(y_{it}|\mathbf{X}_i) &= \int h_t(x_{it},\mathbf{c})\, dG(\mathrm{c}|\mathbf{X}_i) \\ &= \int h_t(x_{it},\mathbf{c})\, dG(\mathbf{c}|\bar{x}_i) \equiv m_t(x_{it},\bar{x}_i) \\ &= \mathrm{E}(y_{it}|x_{it},\bar{x}_i), \qquad t = 1, \ldots, T. \end{aligned} \tag{27}$$

The key is that $\mathrm{E}(y_{it}|\mathbf{X}_i)$ does not depend on $\{x_{i1}, \ldots, x_{iT}\}$ in an unrestricted fashion; it is a function only of $(x_{it},\bar{x}_i)$. If x_{it} were continuous, or took on numerous values, we could use local smoothing 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 [because x_{it} takes on only two values and $\bar{x}_i$ takes on the values $\{0,1/T, \ldots, (T - 1)/T,1\}$]. We can easily write a linear regression function that contains the average treatment effect directly. Iterated expectations implies that the average treatment effect can be computed as $\beta_t = \mathrm{E}[\mathrm{E}[h_t(1,\mathbf{c}_i) - h_t(0,\mathbf{c}_i)|\bar{x}_i]] = \mathrm{E}[m_t(1,\bar{x}_i) - m_t(0,\bar{x}_i)]$. But $m_t(1,\bar{x}_i)$ and $m_t(0,\bar{x}_i)$ take on only $T + 1$ values each. We can characterize these functions simply by defining an exhaustive, mutually exclusive set of dummy indicators: $s_{i1} = 1[\bar{x}_i = 1/T]$, $s_{i2} = 1[\bar{x}_i = 2/T]$, . . . , and $s_{iT} = 1[\bar{x}_i = 1]$. Then $m_t(0,\bar{x}_i) = \alpha_t + \mathbf{s}_i\boldsymbol{\gamma}_t$ and $m_t(1,\bar{x}_i) = \eta_t + \mathbf{s}_i\boldsymbol{\theta}_t$, where $\mathbf{s}_i$ is the $1 \times T$ vector of s_{it}. The average treatment effect is then

$$\beta_t = (\eta_t - \alpha_t) + \mathrm{E}(\mathbf{s}_i)(\boldsymbol{\theta}_t - \boldsymbol{\gamma}_t) \equiv (\eta_t - \alpha_t) + \boldsymbol{\mu}_s\boldsymbol{\delta}_t$$

where $\boldsymbol{\mu}_s \equiv \mathrm{E}(\mathbf{s}_\mathrm{i})$ and $\boldsymbol{\delta}_\mathrm{t} \equiv (\boldsymbol{\theta}_\mathrm{t} - \boldsymbol{\gamma}_\mathrm{t})$. Therefore, we have

$$\begin{aligned} \mathrm{E}(y_{it}|x_{it},\bar{x}_i) &= (1 - x_{it})(\alpha_t + \mathbf{s}_i\boldsymbol{\gamma}_t) + x_{it}(\eta_t + \mathbf{s}_i\boldsymbol{\theta}_t) \\ &= \alpha_t + \mathbf{s}_i\boldsymbol{\gamma}_t + (\eta_t - \alpha_t)x_{it} + x_{it}\mathbf{s}_i\boldsymbol{\delta}_t \\ &= \alpha_t + \mathbf{s}_i\boldsymbol{\gamma}_t + [(\eta_t - \alpha_t) + \boldsymbol{\mu}_s\boldsymbol{\delta}_t]x_{it} + x_{it}(\mathbf{s}_i - \boldsymbol{\mu}_s)\boldsymbol{\delta}_t \\ &= \alpha_t + \mathbf{s}_i\boldsymbol{\gamma}_t + \beta_t x_{it} + x_{it}(\mathbf{s}_i - \boldsymbol{\mu}_s)\boldsymbol{\delta}_t, \quad t = 1, \ldots, T. \end{aligned} \tag{28}$$

Subtracting μ_s from $\mathbf{s}_i$ before forming the T interactions with x_{it} ensures that the coefficient on x_{it} is the ATE.

Equation (4.15) immediately suggests a strategy for estimation β_t. In practice, μ_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

$$y_{it} \text{ on } 1, x_{it}, s_{i1}, \ldots, s_{iT}, x_{it}(s_{i1} - \bar{s}_1), \ldots, x_{it}(s_{iT} - \bar{s}_T), i = 1, \ldots, N, \tag{29}$$

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

If we made the random-effects assumption $D(\mathbf{c}_t|\mathbf{X}_i) = D(\mathbf{c}_i)$, then 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} is the fixed-effects estimate, and so we obtain an estimate of the average treatment effect assumed constant across time. The regression (29) is more flexible than either previous proposal in that it allows ATEs to change over time while allowing $D(\mathbf{c}_i|\bar{x}_i)$ to depend on $\bar{x}_i$ 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 equation (25).

How do the above procedures compare with more common approaches? A general comparison is not possible, because equation (23) puts very little structure on $E(y_{it}|\mathbf{X}_i, \mathbf{c}_i)$ [at the cost of equation (25)]. But suppose y_{it} is a binary response:

$$P(y_{it} = 1|\mathbf{X}_i, c_i) = F(\delta_t + \gamma x_{it} + c_i), \qquad t = 1, \ldots, T, \tag{30}$$

where $F(\cdot)$ is a cumulative distribution function and c_i is a scalar. If we take F to be the logistic function, and the y_{it} are conditionally independent across time, then the conditional maximum likelihood estimator is consistent for γ (and the aggregate time-effect coefficients). Unfortunately, ATEs are not identified, because 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 under the assumption $c_i|\mathbf{X}_i \sim \text{Normal}(\xi_0 + \xi_1 x_{i1} + \cdots + \xi_T x_{iT}, \eta^2)$. (In principle, F could be the logit function, but then implementation of Chamberlain's method is more difficult.) Chamberlain's approach identifies γ as well as the ATEs—see Chamberlain (1984) or Wooldridge (2002, 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 $E(c_i|\mathbf{X}_i)$, at the cost of homoskedasticity and normality in $D(c_i|\mathbf{X}_i)$. The regression procedure outlined above replaces Chamberlain's parametric assumptions with equation (25). The two approaches are complementary in that 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 equation (25).

REFERENCES

Angrist, Joshua D., "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics* 19 (January 2001), 2–16.

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

Chamberlain, Gary, "Panel Data" (pp. 1247–1318), in Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics,* vol. 2 (Amsterdam: North Holland, 1984).

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

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

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

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

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

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

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

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

Wooldridge, Jeffrey M., "Fixed Effects Estimation of the Population-Averaged Slopes in a Panel Data Random Coefficient Model," *Econometric Theory* 19 (April 2003), 411–412.

Wooldridge, Jeffrey M., "Unobserved Heterogeneity and Estimation of Average Partial Effects," in Donald W. K. Andrews and James H. Stock (Eds.), *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg* (Cambridge: Cambridge University Press, 2005).