What's New in Econometrics? NBER, Summer 2007
Lecture 2, Monday, July 30th, 11.00-12.30 am

Linear Panel Data Models

These notes cover some recent topics in linear panel data models. They begin with a "modern" treatment of the basic linear model, and then consider some embellishments, such as random slopes and time-varying factor loads. In addition, fully robust tests for correlated random effects, lack of strict exogeneity, and contemporaneous endogeneity are presented. Section 4 considers estimation of models without strictly exogenous regressors, and Section 5 presents a unified framework for analyzing pseudo panels (constructed from repeated cross sections).

1. Quick Overview of the Basic Model

Most of these notes are concerned with an unobserved effects model defined for a large population. Therefore, we assume random sampling in the cross section dimension. Unless stated otherwise, the asymptotic results are for a fixed number of time periods, *T*, with the number of cross section observations, *N*, getting large.

For some of what we do, it is critical to distinguish the underlying population model of interest and the sampling scheme that generates data that we can use to estimate the population parameters. The standard model can be written, for a generic *i* in the population, as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \ \ t = 1,\ldots,T, \tag{1.1}$$

where η_t is a separate time period intercept (almost always a good idea), $\mathbf{x}_{it}$ is a $1 \times K$ vector of explanatory variables, c_i is the time-constant unobserved effect, and the $\{u_{it} : t = 1,\ldots,T\}$ are idiosyncratic errors. Thanks to Mundlak (1978) and Chamberlain (1982), we view the c_i as random draws along with the observed variables. Then, one of the key issues is whether c_i is correlated with elements of $\mathbf{x}_{it}$.

It probably makes more sense to drop the *i* subscript in (1.1), which would emphasize that the equation holds for an entire population. But (1.1) is useful to emphasizing which factors change only across *t*, which change only change across *i*, and which change across *i* and *t*.It is sometimes convenient to subsume the time dummies in $\mathbf{x}_{it}$.

Ruling out correlation (for now) between u_{it} and $\mathbf{x}_{it}$, a sensible assumption is *contemporaneous exogeneity conditional on* c_i :

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0, t = 1,\ldots,T. \tag{1.2}$$

This equation really defines $\boldsymbol{\beta}$ in the sense that under (1.1) and (1.2),

$$E(y_{it}|\mathbf{x}_{it},c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i, \tag{1.3}$$

so the β_j are partial effects holding fixed the unobserved heterogeneity (and covariates other than x_{tj}).

As is now well known, $\boldsymbol{\beta}$ is not identified only under (1.2). Of course, if we added $Cov(\mathbf{x}_{it},c_i) = \mathbf{0}$ for any t, then $\boldsymbol{\beta}$ is identified and can be consistently estimated by a cross section regression using period t. But usually the whole point is to allow the unobserved effect to be correlated with time-varying $\mathbf{x}_{it}$.

We can allow general correlation if we add the assumption of *strict exogeneity conditional on c_i*:

$$E(u_{it}|\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots,\mathbf{x}_{iT},c_i) = 0, t = 1,\ldots,T, \tag{1.4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},c_i) = E(y_{it}|\mathbf{x}_{it},c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{1.5}$$

If the elements of $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ have suitable time variation, $\boldsymbol{\beta}$ can be consistently estimated by fixed effects (FE) or first differencing (FD), or generalized least squares (GLS) or generalized method of moments (GMM) versions of them. If the simpler methods are used, and even if GLS is used, standard inference can and should be made fully robust to heteroksedasticity and serial dependence that could depend on the regressors (or not). These are the now well-known “cluster” standard errors. With large N and small T, there is little excuse not to compute them.

(Note: Some call (1.4) or (1.5) “strong” exogeneity. But in the Engle, Hendry, and Richard (1983) work, strong exogeneity incorporates assumptions on parameters in different conditional distributions being variation free, and that is not needed here.)

The strict exogeneity assumption is always violated if $\mathbf{x}_{it}$ contains lagged dependent variables, but it can be violated in other cases where $\mathbf{x}_{i,t+1}$ is correlated with u_{it} – a “feedback effect.” An assumption more natural than strict exogeneity is *sequential exogeneity condition on c_i*:

$$E(u_{it}|\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots,\mathbf{x}_{it},c_i) = 0, t = 1,\ldots,T \tag{1.6}$$

or

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{it},c_i) = E(y_{it}|\mathbf{x}_{it},c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{1.7}$$

This allows for lagged dependent variables (in which case it implies that the dynamics in the mean have been completely specified) and, generally, is more natural when we take the view that $\{\mathbf{x}_{it}\}$ might react to shocks that affect y_{it}. Generally, $\boldsymbol{\beta}$ is identified under sequential exogeneity. First differencing and using lags of $\mathbf{x}_{it}$ as instruments, or forward filtering, can be used in simple IV procedures or GMM procedures. (More later.)

If we are willing to assume c_i and $\mathbf{x}_i$ are uncorrelated, then many more possibilities arise (including, of course, identifying coefficients on time-constant explanatory variables). The most convenient way of stating the random effects (RE) assumption is

$$E(c_i|\mathbf{x}_i) = E(c_i), \tag{1.8}$$

although using the linear projection in place of $E(c_i|\mathbf{x}_i)$ suffices for consistency (but usual inference would not generally be valid). Under (1.8), we can used pooled OLS or any GLS procedure, including the usual RE estimator. Fully robust inference is available and should generally be used. (Note: The usual RE variance matrix, which depends only on σ_c^2 and σ_u^2, need not be correctly specified! It still makes sense to use it in estimation but make inference robust.)

It is useful to define two *correlated random effects* assumptions:

$$L(c_i|\mathbf{x}_i) = \psi + \mathbf{x}_i\boldsymbol{\xi}, \tag{1.9}$$

which actually is not an assumption but a definition. For nonlinear models, we will have to actually make assumptions about $D(c_i|\mathbf{x}_i)$, the conditional distribution. Methods based on (1.9) are often said to implement the *Chamberlain device*, after Chamberlain (1982).

Mundlak (1978) used a restricted version, and used a conditional expectation:

$$E(c_i|\mathbf{x}_i) = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \tag{1.10}$$

where $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$. This formulation conserves on degrees of freedom, and extensions are useful for nonlinear models.

If we write $c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i$ or $c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$ and plug into the original equation, for example

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it} \tag{1.11}$$

(absorbing ψ into the time intercepts), then we are tempted to use pooled OLS, or RE estimation because $E(a_i + u_{it}|\mathbf{x}_i) = 0$. Either of these leads to the FE estimator of $\boldsymbol{\beta}$, and to a simple test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$. Later, when we discuss control function methods, it will be handy to run regressions directly that include the time averages. (Somewhat surprisingly, obtain the same algebraic equivalence using Chamberlain's devise. The pooled OLS estimator of $\boldsymbol{\beta}$ is still the FE estimator, even though the $\boldsymbol{\xi}_t$ might change substantially across t.)

Some of us have been pushing for several years the notion that specification tests should be made robust to assumptions that are not directly being tested. (Technically, they should be robust to assumptions that they have no asymptotic power for detecting violations of.) Much progress has been made, but one still sees Hausman statistics computed that maintain a full set of assumptions under the null. Take comparing random effects to fixed effects. The key assumption is (1.8). whether $Var(\mathbf{v}_i|\mathbf{x}_i)$ has the random effects structure, where $v_{it} = c_i + u_{it}$, should not be a critical issue. It makes no sense to report a fully robust variance matrix for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions. (In addition to (1.4) and (1.8), these are $Var(\mathbf{u}_i|\mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T$ and $Var(c_i|\mathbf{x}_i) = Var(c_i)$.) The regression-based Hausman test from (1.11) is very handy for obtaining a fully robust test. More specifically, suppose the model contains a full set of year intercepts as well as time-constant and time-varying explanatory variables:

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}.$$

Now, it is clear that, because we cannot estimate $\boldsymbol{\gamma}$ by FE, it is not part of the Hausman test comparing RE and FE. What is less clear, but also true, is that the coefficients on the time dummies, $\boldsymbol{\eta}$, cannot be included, either. (RE and FE estimation only with aggregate time effects are identical.) In fact, we can only compare the $M \times 1$ estimates of $\boldsymbol{\delta}$, say $\hat{\boldsymbol{\delta}}_{FE}$ and $\hat{\boldsymbol{\delta}}_{RE}$. If we include $\hat{\boldsymbol{\eta}}_{FE}$ and $\hat{\boldsymbol{\eta}}_{RE}$ we introduce a nonsingularity in the asymptotic variance matrix. The regression based test, from the pooled regression

$$y_{it} \text{ on } \mathbf{g}_t, \mathbf{z}_i, \mathbf{w}_{it}, \bar{\mathbf{w}}_i,\ t = 1,\ldots,T;\ i = 1,\ldots,N$$

makes this clear (and that the are M restrictions to test). (Mundlak (1978) suggested this test and Arellano (1993) described the robust version.). Unfortunately, the usual form of the Hausman test does not, and, for example, Stata gets it wrong and tries to include the year dummies in the test (in addition to being nonrobust). The most important problem is that unwarranted degrees of freedom are added to the chi-square distribution, often many extra df, which can produce seriously misleading p-values.

2. New Insights Into Old Estimators

In the past several years, the properties of traditional estimators used for linear models, particularly fixed effects and its instrumental variable counterparts, have been studied under weaker assumptions. We review some of those results here. In these notes, we focus on models without lagged dependent variables or other non-strictly exogenous explanatory variables, although the instrumental variables methods applied to linear models can, in some cases, be applied to models with lagged dependent variables.

2.1. Fixed Effects Estimation in the Correlated Random Slopes Model

The fixed effects (FE) estimator is still the workhorse in empirical studies that employ panel data methods to estimate the effects of time-varying explanatory variables. The attractiveness of the FE estimator is that it allows arbitrary correlation between the additive, unobserved heterogeneity and the explanatory variables. (Pooled methods that do not remove time averages, as well as the random effects (RE) estimator, essentially assume that the unobserved heterogeneity is uncorrelated with the covariates.) Nevertheless, the framework in which the FE estimator is typically analyzed is somewhat restrictive: the heterogeneity is assumed to be additive and is assumed to have a constant coefficients (factor loads) over time. Recently, Wooldridge (2005a) has shown that the FE estimator, and extensions that sweep away unit-specific trends, has robustness properties for estimating the population average effect (PAE) or average partial effect (APE).

We begin with an extension of the usual model to allow for unit-specific slopes,

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{2.1}$$

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

where $\mathbf{b}_i$ is $K \times 1$. Rather than acknowledge that $\mathbf{b}_i$ is unit-specific, we ignore the heterogeneity in the slopes and act as if $\mathbf{b}_i$ is constant for all i. We think c_i might be correlated with at least some elements of $\mathbf{x}_{it}$, and therefore we apply the usual fixed effects estimator. The question we address here is: when does the usual FE estimator consistently estimate the population average effect, $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$.

In addition to assumption (2.2), we naturally need the usual FE rank condition,

$$rank \sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}) = K. \tag{2.3}$$

Write $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$ where the unit-specific deviation from the average, $\mathbf{d}_i$, necessarily has a zero mean. Then

$$y_{it} = c_i + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_{it}\mathbf{d}_i + u_{it} \equiv c_i + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} \tag{2.4}$$

where $v_{it} \equiv \mathbf{x}_{it}\mathbf{d}_i + u_{it}$. A sufficient condition for consistency of the FE estimator along with

(3) is

$$\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{v}_{it}) = \mathbf{0}, t = 1,\ldots,T. \tag{2.5}$$

Along with (2.2), it suffices that $\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\mathbf{d}_i) = \mathbf{0}$ for all t. A sufficient condition, and one that is easier to interpret, is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1,\ldots,T. \tag{2.6}$$

Importantly, condition (2.6) allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. What it rules out is correlation between idiosyncratic movements in $\mathbf{x}_{it}$. We can formalize this statement by writing $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1,\ldots,T$. Then (2.6) holds if $\mathrm{E}(\mathbf{b}_i|\mathbf{r}_{i1},\mathbf{r}_{i2},\ldots,\mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i)$. So $\mathbf{b}_i$ is allowed to be arbitrarily correlated with the permanent component, $\mathbf{f}_i$. (Of course, $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$ is a special representation of the covariates, but it helps to illustrate condition (2.6).) Condition (2.6) is similar in spirit to the Mundlak (1978) assumption applied to the slopes (rather to the intercept):
$\mathrm{E}(\mathbf{b}_i|\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots,\mathbf{x}_{iT}) = \mathrm{E}(\mathbf{b}_i|\bar{\mathbf{x}}_i)$

One implication of these results is that it is a good idea to use a fully robust variance matrix estimator with FE even if one thinks idiosyncratic errors are serially uncorrelated: the term $\ddot{\mathbf{x}}_{it}\mathbf{d}_i$ is left in the error term and causes heteroskedasticity and serial correlation, in general.

These results extend to a more general class of estimators that includes the usual fixed effects and random trend estimator. Write

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it}, \quad t = 1,\ldots,T \tag{2.7}$$

where $\mathbf{w}_t$ is a set of deterministic functions of time. We maintain the standard assumption (2.2) but with $\mathbf{a}_i$ in place of c_i. Now, the "fixed effects" estimator sweeps away $\mathbf{a}_i$ by netting out $\mathbf{w}_t$ from $\mathbf{x}_{it}$. In particular, now let $\ddot{\mathbf{x}}_{it}$ denote the residuals from the regression $\mathbf{x}_{it}$ on
$\mathbf{w}_t, t = 1,\ldots,T$.

In the random trend model, $\mathbf{w}_t = (1,t)$, and so the elements of $\mathbf{x}_{it}$ have unit-specific linear trends removed in addition to a level effect. Removing even more of the heterogeneity from $\{\mathbf{x}_{it}\}$ makes it even more likely that (2.6) holds. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_it + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i,\mathbf{h}_i)$. Of course, individually detrending the $\mathbf{x}_{it}$ requires at least three time periods, and it decreases the variation in $\ddot{\mathbf{x}}_{it}$ compared to the usual FE estimator. Not surprisingly, increasing the dimension of $\mathbf{w}_t$ (subject to the restriction $\dim(\mathbf{w}_t) < T$), generally leads to less precision of the estimator. See Wooldridge (2005a) for further discussion.

Of course, the first differencing transformation can be used in place of, or in conjunction with, unit-specific detrending. For example, if we first difference followed by the within transformation, it is easily seen that a condition sufficient for consistency of the resulting estimator for $\boldsymbol{\beta}$ is

$$\mathrm{E}(\mathbf{b}_i|\Delta\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i),\ \ t = 2,\ldots,T, \tag{2.8}$$

where $\Delta\ddot{\mathbf{x}}_{it} = \Delta\mathbf{x}_{it} - \overline{\Delta\mathbf{x}}$ are the demeaned first differences.

Now consider an important special case of the previous setup, where the regressors that have unit-specific coefficients are time dummies. We can write the model as

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

where, with small T and large N, it makes sense to treat $\{\eta_t : t = 1,\ldots,T\}$ as parameters, like $\boldsymbol{\beta}$. Model (2.9) is attractive because it allows, say, the return to unbobserved "talent" to change over time. Those who estimate, say, firm-level production functions like to allow the importance of unobserved factors, such as managerial skill, to change over time. Estimation of $\boldsymbol{\beta}$, along with the η_t, is a nonlinear problem. What if we just estimate $\boldsymbol{\beta}$ by fixed effects? Let $\mu_c = \mathrm{E}(c_i)$ and write (2.9) as

$$y_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t d_i + u_{it}, t = 1,\ldots,T, \tag{2.10}$$

where $\alpha_t = \eta_t\mu_c$ and $d_i = c_i - \mu_c$ has zero mean In addition, the composite error, $v_{it} \equiv \eta_t d_i + u_{it}$, is uncorrelated with $(\mathbf{x}_{i1},\mathbf{x}_2,\ldots,\mathbf{x}_{iT})$ (as well as having a zero mean). It is easy to see that consistency of the usual FE estimator, which allows for different time period intercepts, is ensured if

$$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, c_i) = \mathbf{0}, t = 1,\ldots,T. \tag{2.11}$$

In other words, the unobserved effects is uncorrelated with the deviations $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$.

If we use the extended FE estimators for random trend models, as above, then we can replace $\ddot{\mathbf{x}}_{it}$ with detrended covariates. Then, c_i can be correlated with underlying levels and trends in $\mathbf{x}_{it}$ (provided we have a sufficient number of time periods).

Using usual FE (with full time period dummies) does not allow us to estimate the η_t, or even determine whether the η_t change over time. Even if we are interested only in $\boldsymbol{\beta}$ when c_i and $\mathbf{x}_{it}$ are allowed to be correlated, being able to detect time-varying factor loads is important because (2.11) is not completely general. It is useful to have a simple test of $\mathrm{H}_0 : \eta_2 = \eta_3 =\ldots= \eta_T$ with some power against the alternative of time-varying coefficients. Then, we can determine whether a more sophisticated estimation method might be needed.

We can obtain a simple variable addition test that can be computed using linear estimation methods if we specify a particular relationship between c_i and $\mathbf{x}_i$. We use the Mundlak (1978) assumption

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i. \tag{2.12}$$

Then

$$y_{it} = \eta_t\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t\bar{\mathbf{x}}_i\boldsymbol{\xi} + \eta_t a_i + u_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \lambda_t\bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + \lambda_t a_i + u_{it}, \tag{2.13}$$

where $\lambda_t = \eta_t - 1$ for all t. Under the null hypothesis, $\lambda_t = 0, t = 2,\ldots,T$. If we impose the null hypothesis, the resulting model is linear, and we can estimate it by pooled OLS of y_{it} on $1, d2_t,\ldots,dT_t, \mathbf{x}_{it}, \bar{\mathbf{x}}_i$ across t and i, where the dr_t are time dummies. A variable addition test that all λ_t are zero can be obtained by applying FE to the equation

$$y_{it} = \alpha_1 + \alpha_2 d2_t + \ldots + \alpha_T dT_t + \mathbf{x}_{it}\boldsymbol{\beta} + \lambda_2 d2_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}) + \ldots + \lambda_T dT_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}) + error_{it}, \tag{2.14}$$

and test the joint significance of the $T-1$ terms $d2_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}),\ldots,dT_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}})$. (The term $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ would drop out of an FE estimation, and so we just omit it.) Note that $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ is a scalar and so the test as $T-1$ degrees of freedom. As always, it is prudent to use a fully robust test (even though, under the null, $\lambda_t a_i$ disappears from the error term).

A few comments about this test are in order. First, although we used the Mundlak device to obtain the test, it does not have to represent the actual linear projection because we are simply adding terms to an FE estimation. Under the null, we do not need to restrict the relationshp between c_i and $\mathbf{x}_i$. Of course, the power of the test may be affected by this choice. Second, the test only makes sense if $\boldsymbol{\xi} \neq 0$; in particular, it cannot be used in a pure random effects environment. Third, a rejection of the null does not necessarily mean that the usual FE estimator is inconsistent for $\boldsymbol{\beta}$: assumption (11) could still hold. In fact, the change in the estimate of $\boldsymbol{\beta}$ when the interaction terms are added can be indicative of whether accounting for time-varying η_t is likely to be important. But, because $\hat{\boldsymbol{\xi}}$ has been estimated under the null, the estimated $\boldsymbol{\beta}$ from (1.14) is not generally consistent.

If we want to estimate the η_t along with $\boldsymbol{\beta}$, we can impose the Mundlak assumption and estimate all parameteres, including $\boldsymbol{\xi}$, by pooled nonlinear regression or some GMM version. Or, we can use Chamberlain’s (1982) less restrictive assumption. But, typically, when we want to allow arbitrary correlation between c_i and $\mathbf{x}_i$, we work directly from (9) and eliminate the c_i. There are several ways to do this. If we maintain that all η_t are different from zero then we can use a quas-differencing method to eliminat c_i. In particular, for $t \geq 2$ we can multiply the $t-1$ equation by η_t/η_{t-1} and subtract the result from the time t equation:

$$\begin{aligned} y_{it} - (\eta_t/\eta_{t-1})y_{i,t-1} &= [\mathbf{x}_{it} - (\eta_t/\eta_{t-1})\mathbf{x}_{i,t-1}]\boldsymbol{\beta} + [\eta_t c_i - (\eta_t/\eta_{t-1})\eta_{t-1}c_i] + [u_{it} - (\eta_t/\eta_{t-1})u_{i,t-1}] \\ &= [\mathbf{x}_{it} - (\eta_t/\eta_{t-1})\mathbf{x}_{i,t-1}]\boldsymbol{\beta} + [u_{it} - (\eta_t/\eta_{t-1})u_{i,t-1}], \quad t \geq 2. \end{aligned}$$

We define $\theta_t = \eta_t/\eta_{t-1}$ and write

$$y_{it} - \theta_t y_{i,t-1} = (\mathbf{x}_{it} - \theta_t \mathbf{x}_{i,t-1})\boldsymbol{\beta} + e_{it}, \; t = 2,\ldots,T, \tag{2.15}$$

where $e_{it} \equiv u_{it} - \theta_t u_{i,t-1}$. Under the strict exogeneity assumption, e_{it} is uncorrelated with every element of $\mathbf{x}_i$, and so we can apply GMM to (2.15) to estimate $\boldsymbol{\beta}$ and $(\theta_2,\ldots,\theta_T)$. Again, this requires using nonlinear GMM methods, and the e_{it} would typically be serially correlated. If we do not impose restrictions on the second moment matrix of $\mathbf{u}_i$, then we would not use any information on the second moments of $\mathbf{e}_i$; we would (eventually) use an unrestricted weighting matrix after an initial estimation.

Using all of $\mathbf{x}_i$ in each time period can result in too many overidentifying restrictions. At time t we might use, say, $\mathbf{z}_{it} = (\mathbf{x}_{it}, \mathbf{x}_{i,t-1})$, and then the instrument matrix $\mathbf{Z}_i$ (with $T-1$ rows) would be $\text{diag}(\mathbf{z}_{i2},\ldots,\mathbf{z}_{iT})$. An initial consistent estimator can be gotten by choosing weighting matrix $(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i)^{-1}$. Then the optimal weighting matrix can be estimated. Ahn, Lee, and Schmidt (2002) provide further discussion.

If $\mathbf{x}_{it}$ contains sequentially but not strictly exogenous explanatory variables – such as a lagged dependent variable – the instruments at time t can only be chosen from $(\mathbf{x}_{i,t-1},\ldots,\mathbf{x}_{i1})$. Holtz-Eakin, Newey, and Rosen (1988) explicitly consider models with lagged dependent variables; more on these models later.

Other transformations can be used. For example, at time $t \geq 2$ we can use the equation

$$\eta_{t-1}y_{it} - \eta_t y_{i,t-1} = (\eta_{t-1}\mathbf{x}_{it} - \eta_t \mathbf{x}_{i,t-1})\boldsymbol{\beta} + e_{it}, \; t = 2,\ldots,T,$$

where now $e_{it} = \eta_{t-1}u_{it} - \eta_t u_{i,t-1}$. This equation has the advantage of allowing $\eta_t = 0$ for some t. The same choices of instruments are available depending on whether $\{\mathbf{x}_{it}\}$ are strictly or sequentially exogenous.

2.2. Fixed Effects IV Estimation with Random Slopes

The results for the fixed effects estimator (in the generalized sense of removing unit-specific means and possibly trends), extend to fixed effects IV methods, provided we add a constant conditional covariance assumption. Murtazashvili and Wooldridge (2007) derive a simple set of sufficient conditions. In the model with general trends, we assume the natural extension of Assumption FEIV.1, that is, $\text{E}(u_{it}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i) = 0$ for all t, along with Assumption FEIV.2. We modify assumption (2.6) in the obvious way: replace $\ddot{\mathbf{x}}_{it}$ with $\ddot{\mathbf{z}}_{it}$, the invididual-specific detrended instruments:

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1,\ldots,T \tag{2.16}$$

But something more is needed. Murtazashvili and Wooldridge (2007) show that, along with the previous assumptions, a sufficient condition is

$$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i), t = 1,\ldots,T. \tag{2.17}$$

Note that the covariance $\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i)$, a $K \times K$ matrix, need not be zero, or even constant across time. In other words, we can allow the detrended covariates to be arbitrarily correlated with the heterogeneous slopes, and that correlation can change in any way across time. But the *conditional* covariance cannot depend on the time-demeaned instruments. (This is an example of how it is important to distinguish between a conditional expectation and an unconditional one: the implicit error in the equation generally has an unconditional mean that changes with *t*, but its conditional mean does not depend on $\ddot{\mathbf{z}}_{it}$, and so using $\ddot{\mathbf{z}}_{it}$ as IVs is valid provided we allow for a full set of dummies.) Condition (2.17) extends to the panel data case the assumption used by Wooldridge (2003a) in the cross section case.

We can easily show why (2.17) suffices with the previous assumptions. First, if $\mathrm{E}(\mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathbf{0}$ – which follows from $\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\mathbf{b}_i)$ – then $\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i'|\ddot{\mathbf{z}}_{it})$, and so $\mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i) \equiv \gamma_t$ under the previous assumptions. Write $\ddot{\mathbf{x}}_{it}\mathbf{d}_i = \gamma_t + r_{it}$ where $\mathrm{E}(r_{iti}|\ddot{\mathbf{z}}_{it}) = 0, t = 1,\ldots,T$. Then we can write the transformed equation as

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{\mathbf{x}}_{it}\mathbf{d}_i + \ddot{u}_{it} = \ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \gamma_t + r_{it} + \ddot{u}_{it}. \tag{2.18}$$

Now, if $\mathbf{x}_{it}$ contains a full set of time period dummies, then we can absorb γ_t into $\ddot{\mathbf{x}}_{it}$, and we assume that here. Then the sufficient condition for consistency of IV estimators applied to the transformed equations is $\mathrm{E}[\ddot{\mathbf{z}}_{it}'(r_{it} + \ddot{u}_{it})] \;= \mathbf{0}$,.and this condition is met under the maintained assumptions. In other words, under (2.16) and (2.17), the fixed effects 2SLS estimator is consistent for the average population effect, $\boldsymbol{\beta}$. (Remember, we use “fixed effects” here in the general sense of eliminating the unit-specific trends, $\mathbf{a}_i$.) We must remember to include a full set of time period dummies if we want to apply this robustness result, something that should be done in any case. Naturally, we can also use GMM to obtain a more efficient estimator. If $\mathbf{b}_i$ truly depends on *i*, then the composite error $r_{it} + \ddot{u}_{it}$ is likely serially correlated and heteroskedastic. See Murtazashvili and Wooldridge (2007) for further discussion and similation results on the peformance of the FE2SLS estimator. They also provide examples where the key assumptions cannot be expected to hold, such as when endogenous elements of

$\mathbf{x}_{it}$ are discrete.

3. Behavior of Estimators without Strict Exogeneity

As is well known, both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the conditional strict exogeneity assumption. But it is also pretty well known that, at least under certain assumptions, the FE estimator can be expected to have less "bias" (actually, inconsistency) for larger T. One assumption is contemporaneous exogeneity, (1.2). If we maintain this assumption, assume that the data series $\{(\mathbf{x}_{it}, u_{it}) : t = 1,\ldots,T\}$ is "weakly dependent" – in time series parlance, integrated of order zero, or I(0) – then we can show that

$$\text{plim } \hat{\boldsymbol{\beta}}_{FE} = \boldsymbol{\beta} + O(T^{-1}) \tag{3.1}$$

$$\text{plim } \hat{\boldsymbol{\beta}}_{FD} = \boldsymbol{\beta} + O(1). \tag{3.2}$$

In some special cases – the AR(1) model without extra covariates – the "bias" terms can be calculated. But not generally. The FE (within) estimator averages across T, and this tends to reduce the bias.

Interestingly, the same results can be shown if $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ has unit roots as long as $\{u_{it}\}$ is I(0) and contemporaneous exogeneity holds. But there is a catch: if $\{u_{it}\}$ is I(1) – so that the time series version of the "model" would be a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not cointegrated), then (3.1) is no longer true. And, of course, the first differencing means any unit roots are eliminated. So, once we start appealing to "large T" to prefer FE over FD, we must start being aware of the time series properties of the series.

The same comments hold for IV versions of the estimators. Provided the instruments are contemporaneously exogenous, the FEIV estimator has bias of order T^{-1}, while the bias in the FDIV estimator does not shrink with T. The same caveats about applications to unit root processes also apply.

Because failure of strict exogeneity causes inconsistency in both FE and FD estimation, it is useful to have simple tests. One possibility is to obtain a Hausman test directly comparing the FE and FD estimators. This is a bit cumbersome because, when aggregate time effects are included, the difference in the estimators has a singular asymptotic variance. Plus, it is somewhat difficult to make the test fully robust.

Instead, simple regression-based strategies are available. Let $\mathbf{w}_{it}$ be the $1 \times Q$ vector, a subset of $\mathbf{x}_{it}$ suspected of failing strict exogeneity. A simple test of strict exogeneity, specifically looking for feedback problems, is based on

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it},\ t = 1,\ldots,T-1. \tag{3.3}$$

Estimate the equation by fixed effects and test $H_0 : \boldsymbol{\delta} = \mathbf{0}$ (using a fully robust test). Of course, the test may have little power for detecting contemporaneous endogeneity.

In the context of FEIV we can test whether a subset of instruments fails strict exogeneity by writing

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{h}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it},\ t = 1,\ldots,T-1, \tag{3.4}$$

where $\mathbf{h}_{it}$ is a subset of the instruments, $\mathbf{z}_{it}$. Now, estimate the equation by FEIV using instruments $(\mathbf{z}_{it}, \mathbf{h}_{i,t+1})$ and test coefficients on the latter.

It is also easy to test for contemporaneous endogeneity of certain regressors, even if we allow some regressors to be endogenous under the null. Write the model now as

$$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + \mathbf{y}_{it3}\boldsymbol{\gamma}_1 + c_{i1} + u_{it1}, \tag{3.5}$$

where, in an FE environment, we want to test $\mathrm{H}_0 : \mathrm{E}(\mathbf{y}_{it3}' u_{it1}) = \mathbf{0}$. Actually, because we are using the within transformation, we are really testing strict exogeneity of $\mathbf{y}_{it3}$, but we allow all variables to be correlated with c_{i1}. The variables $\mathbf{y}_{it2}$ are allowed to be endogenous under the null – provided, of course, that we have sufficient instruments excluded from the structural equation that are uncorrelated with u_{it1} in every time period. We can write a set of reduced forms for elements of $\mathbf{y}_{it3}$ as

$$\mathbf{y}_{it3} = \mathbf{z}_{it}\boldsymbol{\Pi}_3 + \mathbf{c}_{i3} + \mathbf{v}_{it3}, \tag{3.6}$$

and obtain the FE residuals, $\hat{\ddot{\mathbf{v}}}_{it3} = \ddot{\mathbf{y}}_{it3} - \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Pi}}_3$, where the columns of $\hat{\boldsymbol{\Pi}}_3$ are the FE estimates of the reduced forms, and the double dots denotes time-demeaning, as usual. Then, estimate the equation

$$\ddot{y}_{it1} = \ddot{\mathbf{z}}_{it1}\boldsymbol{\delta}_1 + \ddot{\mathbf{y}}_{it2}\boldsymbol{\alpha}_1 + \ddot{\mathbf{y}}_{it3}\boldsymbol{\gamma}_1 + \hat{\ddot{\mathbf{v}}}_{it3}\boldsymbol{\rho}_1 + error_{it1} \tag{3.7}$$

by pooled IV, using instruments $(\ddot{\mathbf{z}}_{it}, \ddot{\mathbf{y}}_{it3}, \hat{\ddot{\mathbf{v}}}_{it3})$. The test of the null that $\mathbf{y}_{it3}$ is exogenous is just the (robust) test that $\boldsymbol{\rho}_1 = \mathbf{0}$, and the usual robust test is valid with adjusting for the first-step estimation.

An equivalent approach is to define $\hat{\mathbf{v}}_{it3} = \mathbf{y}_{it3} - \mathbf{z}_{it}\hat{\boldsymbol{\Pi}}_3,$ where $\hat{\boldsymbol{\Pi}}_3$ is still the matrix of FE coefficients, add these to equation (3.5), and apply FE-IV, using a fully robust test. Using a built-in command can lead to problems because the test is rarely made robust and the degrees of freedom are often incorrectly counted.

4. Instrumental Variables Estimation under Sequential Exogeneity

We now consider IV estimation of the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it},\ t = 1,\ldots,T, \tag{4.1}$$

under sequential exogeneity assumptions. Some authors simply use

$$E(\mathbf{x}_{is}'u_{it}) = 0, s = 1,\ldots,T, t = 1,\ldots,T. \tag{4.2}$$

As always, $\mathbf{x}_{it}$ probably includes a full set of time period dummies. This leads to simple moment conditions after first differencing:

$$\mathrm{E}(\mathbf{x}_{is}'\Delta u_{it}) = \mathbf{0}, s = 1,\ldots,t-1;\ t = 2,\ldots,T. \tag{4.3}$$

Therefore, at time t, the available instruments in the FD equation are in the vector $\mathbf{x}_{i,t-1}^{o}$, where

$$\mathbf{x}_{it}^{o} \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2},\ldots,\mathbf{x}_{it}). \tag{4.4}$$

Therefore, the matrix of instruments is simply

$$\mathbf{W}_i = \mathrm{diag}(\mathbf{x}_{i1}^{o}, \mathbf{x}_{i2}^{o},\ldots,\mathbf{x}_{i,T-1}^{o}), \tag{4.5}$$

which has $T-1$ rows. Because of sequential exogeneity, the number of valid instruments increases with t.

Given $\mathbf{W}_i$, it is routine to apply GMM estimation. But some simpler strategies are available that can be used for comparison or as the first-stage estimator in computing the optimal weighting matrix. One useful one is to estimate a reduced form for $\Delta\mathbf{x}_{it}$ separately for each t. So, at time t, run the regression $\Delta\mathbf{x}_{it}$ on $\mathbf{x}_{i,t-1}^{o}$, $i = 1,\ldots,N$, and obtain the fitted values, $\widehat{\Delta\mathbf{x}}_{it}$. Of course, the fitted values are all $1 \times K$ vectors for each t, even though the number of available instruments grows with t. Then, estimate the FD equation

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it},\ t = 2,\ldots,T \tag{4.6}$$

by pooled IV using instruments (not regressors) $\widehat{\Delta\mathbf{x}}_{it}$. It is simple to obtain robust standard errors and test statistics from such a procedure because the first stage estimation to obtain the instruments can be ignored (asymptotically, of course).

One potential problem with estimating the FD equation by IVs that are simply lags of $\mathbf{x}_{it}$ is that changes in variables over time are often difficult to predict. In other words, $\Delta\mathbf{x}_{it}$ might have little correlation with $\mathbf{x}_{i,t-1}^{o}$, in which case we face a problem of weak instruments. In one case, we even lose identification: if $\mathbf{x}_{it} = \boldsymbol{\lambda}_t + \mathbf{x}_{i,t-1} + \mathbf{e}_{it}$ where $\mathrm{E}(\mathbf{e}_{it}|\mathbf{x}_{i,t-1},\ldots,\mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of $\mathbf{x}_{it}$ are random walks with drift – then $\mathrm{E}(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1},\ldots,\mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails.

If we impose what is generally a stronger assumption, **dynamic completeness in the conditional mean**,

$$\mathrm{E}(u_{it}|\mathbf{x}_{it}, y_{i,t-1}\mathbf{x}_{i,t-1},\ldots,y_{i1},\mathbf{x}_{i1},c_i) = 0, \quad t = 1,\ldots,T, \tag{4.7}$$

then more moment conditions are available. While (4.7) implies that virtually any nonlinear function of the $\mathbf{x}_{it}$ can be used as instruments, the focus has been only on zero covariance assumptions (or (4.7) is stated as a linear projection). The key is that (4.7) implies that $\{u_{it} : t = 1,\ldots,T\}$ is a serially uncorrelated sequence and u_{it} is uncorrelated with c_i for all t. If we use these facts, we obtain moment conditions first proposed by Ahn and Schmidt (1995) in the context of the AR(1) unosberved effects model; see also Arellano and Honoré (2001). They can be written generally as

$$E[(\Delta y_{i,t-1} - \Delta\mathbf{x}_{i,t-1}\boldsymbol{\beta})'(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, t = 3,\ldots,T. \tag{4.8}$$

Why do these hold? Because all u_{it} are uncorrelated with c_i, and $\{u_{i,t-1},\ldots,u_{i1}\}$ are uncorrelated with $c_i + u_{it}$. So $(u_{i,t-1} - u_{i,t-2})$ is uncorrelated with $(c_i + u_{it})$, and the resulting moment conditions can be written in terms of the parameters as (4.8). Therefore, under (4.7), we can add the conditions (4.8) to (4.3) to improve efficiency – in some cases quite substantially with persistent data.

Of course, we do not always intend for models to be dynamically complete in the sense of (4.7). Often, we estimate static models or finite distributed lag models – that is, models without lagged dependent variables – that have serially correlated idiosyncratic errors, and the explanatory variables are not strictly exogenous and so GLS procedures are inconsistent. Plus, the conditions in (4.8) are nonlinear in parameters.

Arellano and Bover (1995) suggested instead the restrictions

$$Cov(\Delta\mathbf{x}_{it}', c_i) = 0, \; t = 2,\ldots,T. \tag{4.9}$$

Interestingly, this is zero correlation, FD version of the conditions from Section 2 that imply we can ignore heterogeneous coefficients in estimation under strict exogeneity. Under (4.9), we have the moment conditions from the levels equation:

$$E[\Delta\mathbf{x}_{it}'(y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, t = 2,\ldots,T, \tag{4.10}$$

because $y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} = c_i + u_{it}$ and u_{it} is uncorrelated with $\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$. We add an intercept, α, explicitly to the equation to allow a nonzero mean for c_i. Blundell and Bond (1999) apply these moment conditions, along with the usual conditions in (4.3), to estimate firm-level production functions. Because of persistence in the data, they find the moments in (4.3) are not especially informative for estimating the parameters. Of course, (4.9) is an extra set of assumptions.

The previous discussion can be applied to the AR(1) model, which has received much attention. In its simplest form we have

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}, t = 1,\ldots,T, \tag{4.11}$$

so that, by convention, our first observation on y is at $t = 0$. Typically the minimal assumptions imposed are

$$E(y_{is}u_{it}) = 0,\ s = 0,\ldots,t-1,\ t = 1,\ldots,T, \tag{4.12}$$

in which case the available instruments at time t are $\mathbf{w}_{it} = (y_{i0},\ldots,y_{i,t-2})$ in the FD equation

$$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta u_{it}, t = 2,\ldots,T. \tag{4.13}$$

In oher words, we can use

$$E[y_{is}(\Delta y_{it} - \rho\Delta y_{i,t-1}) = 0,\ s = 0,\ldots,t-2,\ t = 2,\ldots,T. \tag{4.14}$$

Anderson and Hsiao (1982) proposed pooled IV estimation of the FD equation with the single instrument $y_{i,t-2}$ (in which case all $T-1$ periods can be used) or $\Delta y_{i,t-2}$ (in which case only $T-2$ periods can be used). We can use pooled IV where $T-1$ separate reduced forms are estimated for $\Delta y_{i,t-1}$ as a linear function of $(y_{i0},\ldots,y_{i,t-2})$. The fitted values $\widehat{\Delta y}_{i,t-1}$, can be used as the instruments in (4.13) in a pooled IV estimation. Of course, standard errors and inference should be made robust to the MA(1) serial correlation in Δu_{it}. Arellano and Bond (1991) suggested full GMM estimation using all of the available instruments $(y_{i0},\ldots,y_{i,t-2})$, and this estimator uses the conditions in (4.12) efficiently.

Under the dynamic completeness assumption

$$\mathrm{E}(u_{it}|y_{i,t-1},y_{i,t-2},\ldots,y_{i0},c_i) = 0, \tag{4.15}$$

the Ahn-Schmidt extra moment conditions in (4.8) become

$$E[(\Delta y_{i,t-1} - \rho\Delta y_{i,t-2})(y_{it} - \rho y_{i,t-1})] = 0, t = 3,\ldots,T. \tag{4.16}$$

Blundell and Bond (1998) noted that if the condition

$$Cov(\Delta y_{i1},c_i) = Cov(y_{i1} - y_{i0},c_i) = 0 \tag{4.17}$$

is added to (4.15) then the combinded set of moment conditions becomes

$$E[\Delta y_{i,t-1}(y_{it} - \alpha - \rho y_{i,t-1})] = 0, t = 2,\ldots,T, \tag{4.18}$$

which can be added to the usual moment conditions (4.14). Therefore, we have two sets of moments linear in the parameters. The first, (4.14), use the differenced equation while the second, (4.18), use the levels. Arellano and Bover (1995) analyzed GMM estimators from these equations generally.

As discussed by Blundell and Bond (1998), condition (4.17) can be intepreted as a restriction on the initial condition, y_{i0}. To see why, write $y_{i1} - y_{i0} = \rho y_{i0} + c_i + u_{i1} - y_{i0} = (1-\rho)y_{i0} + c_i + u_{i1}$. Because u_{i1} is uncorrelated with c_i, (4.17) becomes

$$Cov((1-\rho)y_{i0} + c_i, c_i) = 0. \tag{4.19}$$

Write y_{i0} as a deviation from its steady state, $c_i/(1-\rho)$ (obtained for $|\rho| < 1$ be recursive subsitution and then taking the limit), as

$$y_{i0} = c_i/(1-\rho) + r_{i0}. \tag{4.20}$$

Then $(1-\rho)y_{i0} + c_i = (1-\rho)r_{i0}$, and so (4.17) reduces to

$$Cov(r_{i0}, c_i) = 0. \tag{4.21}$$

In other words, the deviation of y_{i0} from its steady state is uncorrelated with the steady state. Blundell and Bond (1998) contains discussion of when this condition is reasonable. Of course, it is not for $\rho = 1$, and it may not be for ρ “close” to one. On the other hand, as shown by Blundell and Bond (1998), this restriction, along with the Ahn-Schmidt conditions, is very informative for ρ close to one. Hahn (1999) shows theoretically that such restrictions can greatly increase the information about ρ.

The Ahn-Schmidt conditions (4.16) are attractive in that they are implied by the most natural statement of the model, but they are nonlinear and therefore more difficult to use. By adding the restriction on the initial condition, the extra moment condition also means that the full set of moment conditions is linear. Plus, this approach extends to general models with only sequentially exogenous variabes as in (4.10). Extra moment assumptions based on homoskedasticity assumptions – either conditional or unconditional – have not been used nearly as much, probably because they impose conditions that have little if anything to do with the economic hypotheses being tested.

Other approaches to dynamic models are based on maximum likelihood estimation or generalized least squares estimation of a particular set of conditional means. Approaches that condition on the initial condition y_{i0}, an approach suggested by Chamberlain (1980), Blundell and Smith (1991), and Blundell and Bond (1998), seem especially attractive. For example, suppose we assume that

$$D(y_{it}|y_{i,t-1},y_{i,t-2},\ldots,y_{i1},y_{i0},c_i) = \text{Normal}(\rho y_{i,t-1}+c_i,\sigma_u^2),\ t=1,2,\ldots,T.$$

Then the distribution of $(y_{i1},\ldots,y_{iT})$ given $(y_{i0}=y_0,c_i=c)$ is just the product of the normal distributions:

$$\prod_{t=1}^{T}\sigma_u^{-T}\phi[(y_t-\rho y_{t-1}-c)/\sigma_u].$$

We can obtain a usable density for (conditional) MLE by assuming

$$c_i|y_{i0} \sim\text{Normal}(\varphi_0+\xi_0 y_{i0},\sigma_a^2).$$

The log likelihood function is obtained by taking the log of

$$\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T}(1/\sigma_u)^T\phi[(y_{it}-\rho y_{i,t-1}-c)/\sigma_u].\right)(1/\sigma_a)\phi[(c-\varphi_0-\xi_0 y_{i0})/\sigma_a)dc.$$

Of course, if this is this represents the correct density of $(y_{i1},\ldots,y_{iT})$ given y_{i0} then the MLE is consistent and $\sqrt{N}$-asymptotically normal (and efficient among estimators that condition on y_{i0}).

A more robust approach is to use a generalized least squares approach where $E(\mathbf{y}_i|y_{i0})$ and $Var(\mathbf{y}_i|y_{i0})$ are obtained, and where the latter could even be misspecified. Like with the MLE approach, this results in estimation that is highly nonlinear in the parameters and is used less often than the GMM procedures with linear moment conditions.

The same kinds of moment conditions can be used in extensions of the AR(1) model, such as

$$y_{it} = \rho y_{i,t-1}+\mathbf{z}_{it}\boldsymbol{\gamma}+c_i+u_{it},\ \ t=1,\ldots,T.$$

If we difference to remove c_i, we can then use exogeneity assumptions to choose instruments. The FD equation is

$$\Delta y_{it} = \rho\Delta y_{i,t-1}+\Delta\mathbf{z}_{it}\boldsymbol{\gamma}+\Delta u_{it},\ \ t=1,\ldots,T,$$

and if the $\mathbf{z}_{it}$ are strictly exogenous with respect to $\{u_{i1},\ldots,u_{iT}\}$ then the available instruments (in addition to time period dummies) are $(\mathbf{z}_i,y_{i,t-2},\ldots,y_{i0})$. We might not want to use all of $\mathbf{z}_i$ for every time period. Certainly we would use $\Delta\mathbf{z}_{it}$, and perhaps a lag, $\Delta\mathbf{z}_{i,t-1}$. If we add sequentially exogenous variables, say $\mathbf{h}_{it}$, to (11.62) then $(\mathbf{h}_{i,t-1},\ldots,\mathbf{h}_{i1})$ would be added to the list of instruments (and $\Delta\mathbf{h}_{it}$ would appear in the equation). We might also add the Arellano and Bover conditions (4.10), or at least the Ahn and Schmidt conditions (4.8).

As a simple example of methods for dynamic models, consider a dynamic air fare equation for routes in the United States:

$$lfare_{it} = \theta_t + \rho\, lfare_{i,t-1} + \gamma\, concen_{it} + c_i + u_{it},$$

where we include a full set of year dummies. We assume the concentration ratio, $concen_{it}$, is strictly exogenous and that at most one lag of *lfare* is needed to capture the dynamics. The data are for 1997 through 2000, so the equation is specified for three years. After differencing, we have only two years of data:

$$\Delta lfare_{it} = \eta_t + \rho\Delta lfare_{i,t-1} + \gamma\Delta concen_{it} + \Delta u_{it},\ t = 1999, 2000.$$

If we estimate this equation by pooled OLS, the estimators are inconsistent because $\Delta lfare_{i,t-1}$ is correlated with Δu_{it}; we include the OLS estimates for comparison. We apply the simple pooled IV procedure, where separate reduced forms are estimated for $\Delta lfare_{i,t-1}$: one for 1999, with $lfare_{i,t-2}$ and $\Delta concen_{it}$ in the reduced form, and one for 2000, with $lfare_{i,t-2}$, $lfare_{imt-3}$ and $\Delta concen_{it}$ in the reduced form. The fitted values are used in the pooled IV estimation, with robust standard errors. (We only use $\Delta concen_{it}$ in the IV list at time *t*.) Finally, we apply the Arellano and Bond (1991) GMM procedure.

Dependent Variable:	*lfare*		
	(1)	(2)	(3)
Explanatory Variable	Pooled OLS	Pooled IV	Arellano-Bond
$lfare_{-1}$	$-.126$	$.219$	$.333$
	$(.027)$	$(.062)$	$(.055)$
concen	$.076$	$.126$	$.152$
	$(.053)$	$(.056)$	$(.040)$
N	$1,149$	$1,149$	$1,149$

As is seen from column (1), the pooled OLS estimate of ρ is actually negative and statistically different from zero. By contrast, the two IV methods give positive and statistically significant estimates. The GMM estimate of ρ is larger, and it also has a smaller standard error (as we would hope for GMM).

The previous example has small *T*, but some panel data applications have reasonable large *T*. Arellano and Alvarez (1998) show that the GMM estimator that accounts for the MA(1) serial correlation in the FD errors has desirable properties when *T* and *N* are both large, while the pooled IV estimator is actually inconsistent under asymptotics where $T/N \to a > 0$. See Arellano (2003, Chapter 6) for discussion.

5. Pseudo Panels from Pooled Cross Sections

In cases where panel data sets are not available, we can still estimate parameters in an underlying panel population model if we can obtain random samples in different periods. Many surveys are done annually by obtaining a different random (or stratified) sample for each year. Deaton (1985) showed how to identify and estimate parameters in panel data models from pooled cross sections. As we will see, however, identification of the parameterse can be tenuous.

Deaton (1985) was careful about distinguishing between the population model on the one hand and the sampling scheme on the other. This distinction is critical for understanding the nature of the identification problem, and in deciding the appropriate asymptotic analysis. The recent literature has tended to write “models” at the cohort or group level, which is not in the spirit of Deaton’s original work. (Angrist (1991) actually has panel data, but uses averages in each t to estimate parameters of a labor supply function.)

In what follows, we are interested in estimating the parameters of the population model

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f + u_t,\ t = 1,\ldots,T, \tag{5.1}$$

which is best viewed as representing a population defined over T time periods. For this setup to make sense, it must be the case that we can think of a stationary population, so that the same units are represented in each time period. Because we allow a full set of period intercepts, $E(f)$ is never separately identified, and so we might as well set it to zero.

The random quantities in (5.1) are the response variable, y_t, the covariates, $\mathbf{x}_t$ (a $1 \times K$ vector), the unobserved effect, f, and the unobserved idiosyncratic errors, $\{u_t : t = 1,\ldots,T\}$. Like our previous analysis, we are thinking of applications with a small number of time periods, and so we view the intercepts, η_t, as parameters to estimate, along with the $K \times 1$ vector parameter – which is ultimately of interest. We consider the case where all elements of $\mathbf{x}_t$ have some time variation.

As it turns out, to use the standard analysis, we do not even have to assume contemporaneous exogeneity conditional on f, that is,

$$E(u_t|\mathbf{x}_t, f) = 0, t = 1,\ldots,T, \tag{5.2}$$

although this is a good starting point to determine reasonable population assumptions.

Naturally, iterated expectations implies

$$E(u_t|f) = 0, t = 1,\ldots,T, \tag{5.3}$$

and (5.3) is sensible in the context of (5.1). Unless stated otherwise, we take it to be true. Because f aggregates all time-constant unobservables, we should think of (5.3) as implying that $E(u_t|g) = 0$ for any time-constant variable g, whether unobserved or observed. In other words, in the leading case we should think of (5.1) as representing $E(y_t|\mathbf{x}_t, f)$ where any time constant factors are lumped into f.

With a (balanced) panel data set, we would have a random sample in the cross section. Therefore, for a random draw i, $\{(\mathbf{x}_{it}, y_{it}), t = 1,\ldots,T\}$, we would then write the model as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + f_i + u_{it},\ t = 1,\ldots,T. \tag{5.4}$$

While this notation can cause confusion later when we sample from each cross section, it has the benefit of explictly labelling quantities as changing only across t, changing only across i, or changing across both.

The idea of using independent cross sections to estimate parameters from panel data models is based on a simple insight of Deaton’s. Assume that the population for which (5.1) holds is divided into G groups (or cohorts). This designation cannot depend on time. For example, it is common to birth year to define the groups, or even ranges of birth year. For a random draw i satisfying (5.4), let g_i be the group indicator, taking on a value in $\{1,2,\ldots,G\}$. Then, by our earlier discussion,

$$E(u_{it}|g_i) = 0, t = 1,\ldots,T, \tag{5.5}$$

essentially by definition. In other words, the η_t account for any change in the average unobservables over time and f_i accounts for any time-constant factors.

Taking the expected value of (5.4) conditional on group membership and using only (5.5), we have

$$E(y_{it}|g_i = g) = \eta_t + E(\mathbf{x}_{it}|g_i = g)\boldsymbol{\beta} + E(f_i|g_i = g),\ t = 1,\ldots,T. \tag{5.6}$$

Again, this expession represents an underlying population, but where we have partitioned the population into G groups.

Several authors after Deaton, including Collado (1997) and Verbeek and Vella (2005), have left $E(u_{it}|g_i = g)$ as part of the “error term,” with the notation $u_{gt}^* = E(u_{it}|g_i = g)$. In fact, these authors have criticized previous work by Moffitt (1993) for making the “asssumption” that $u_{gt}^* = 0$. But, as Deaton showed, if we start with the underlying population model (5.1), then $E(u_{it}|g_i = g) = 0$ for all g follows directly. Nevertheless, as we will discuss later, the key assumption is that the structural model (5.1) does not require a full set of group/time effects. If such effects are required, then one way to think about the resulting misspecification is that $E(u_{it}|g_i = g)$ is not zero.

If we define the population means

$$\begin{aligned} \alpha_g &= E(f_i|g_i = g) \\ \mu^y_{gt} &= E(y_{it}|g_i = g) \\ \boldsymbol{\mu}^{\mathbf{x}}_{gt} &= E(\mathbf{x}_{it}|g_i = g) \end{aligned} \tag{5.7}$$

for $g = 1,\ldots,G$ and $t = 1,\ldots,T$ we have

$$\mu^y_{gt} = \eta_t + \boldsymbol{\mu}^{\mathbf{x}}_{gt}\boldsymbol{\beta} + \alpha_g,\ g = 1,\ldots,G,\ t = 1,\ldots,T. \tag{5.8}$$

(Many authors use the notation y^*_{gt} in place of μ^y_{gt}, and similarly for $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$, but, at this point, such a notation gives the wrong impression that the means defined in (5.7) are random variables. They are not. They are group/time means defined on the underlying population.)

Equation (5.8) is remarkable in that it holds without any assumptions restricting the dependence between $\mathbf{x}_{it}$ and u_{ir} across t and r. In fact, $\mathbf{x}_{it}$ can contain lagged dependent variables, most commonly $y_{i,t-1}$, or explanatory variables that are contemporaneously endogenous (as occurs under measurement error in the original population model, an issue that was important to Angrist (1991)). This probably should make us a little suspicious, as the problems of lagged dependent variable, measurement error, and other violations of strict exogeneity are tricky to handle with true panel data.

(In estimation, we will deal with the fact that there are not really $T + G$ parameters in η_t and α_g to estimate; there are only $T + G - 1$. The lost degree of freedom comes from $E(f) = 0$, which puts a restriction on the α_g. With the groups of the same size in the population, the restriction is that the α_g sum to zero.)

If we take (5.8) as the starting point for estimating $\boldsymbol{\beta}$ (along with η_t and α_g), then the issues become fairly clear. If we have sufficient observations in the group/time cells, then the means μ^y_{gt} and $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ can be estimated fairly precisly, and these can be used in a minimum distance estimation framework to estimate $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ consists of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$ (where, say, we set $\eta_1 = 0$ as the normalization).

Before discussing estimation details, it is useful to study (5.8) in more detail to determine some simple, and common, strategies. Because (5.8) looks itself like a panel data regression equation, methods such as "OLS," "fixed effects," and "first differencing" have been applied to sample averages. It is informative to apply these to the population. First suppose that we set each α_g to zero and set all of the time intercepts, η_t, to zero. For notational simplicity, we also drop an overall "intercept," but that would be included at a minimum. Then $\mu_{gt}^{y} = \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta}$ and if we premultiply by $\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}$, average across g and t, and then assume we can invert $\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\boldsymbol{\mu}_{gt}^{\mathbf{x}}$, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\boldsymbol{\mu}_{gt}^{\mathbf{x}}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\mu_{gt}^{y}\right). \tag{5.9}$$

This means that the population parameter, $\boldsymbol{\beta}$, can be written as a pooled OLS regression of the population group/time means μ_{gt}^{y} on the group/time means $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. Naturally, if we have "good" estimates of these means, then it will make sense to estimate $\boldsymbol{\beta}$ by using the same regression on the sample means. But, so far, this is all in the population. We can think of (5.9) as the basis for a method of moments procedure. It is important that we treat $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ and μ_{gt}^{y} symetrically, that is, as population means to be estimated, whether the $\mathbf{x}_{it}$ are strictly, sequentially, or contemporaneous exogenous – or none of these – in the original model.

When we allow different group means for f_i, as seems critical, and different time period intercepts, which also is necessary for a convincing analysis, we can easily write $\boldsymbol{\beta}$ as an "OLS" estimator by subtracting of time and group averages. While we cannot claim that these expressions will result in efficient estimators, they can shed light on whether we can expect (5.8) to lead to precise estimation of $\boldsymbol{\beta}$. First, without separate time intercepts we have

$$\mu_{gt}^{y} - \bar{\mu}_{g}^{y} = (\boldsymbol{\mu}_{gt}^{\mathbf{x}} - \bar{\boldsymbol{\mu}}_{g}^{\mathbf{x}})\boldsymbol{\beta},\ g = 1,\ldots,G,;\ t = 1,\ldots,T, \tag{5.10}$$

where the notation should be clear, and then one expression for $\boldsymbol{\beta}$ is (5.9) but with $\boldsymbol{\mu}_{gt}^{\mathbf{x}} - \bar{\boldsymbol{\mu}}_{g}^{\mathbf{x}}$ in place of $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. Of course, this makes it clear that identification of $\boldsymbol{\beta}$ more difficult when the α_g are allowed to differ. Further, if we add in the year intercepts, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\mu_{gt}^{y}\right) \tag{5.11}$$

where $\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$ is the vector of residuals from the pooled regression

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} \text{ on } 1,\ d2,\ldots,dT,\ c2,\ \ldots,\ cG, \tag{5.12}$$

where *dt* denotes a dummy for period *t* and *cg* is a dummy variable for group *g*.

There are other expressions for $\boldsymbol{\beta}$, too. (Because $\boldsymbol{\beta}$ is generally overidentified, there are many ways to write it in terms of the population moments. For example, if we difference and then take away group averages, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=2}^{T} \Delta \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \Delta \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=2}^{T} \Delta \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \Delta \mu_{gt}^{y} \right) \tag{5.13}$$

where $\Delta \boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\mu}_{gt}^{\mathbf{x}} - \boldsymbol{\mu}_{g,t-1}^{\mathbf{x}}$ and $\Delta \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} = \Delta \boldsymbol{\mu}_{gt}^{\mathbf{x}} - G^{-1} \sum_{h=1}^{G} \Delta \boldsymbol{\mu}_{ht}^{\mathbf{x}}$.

Equations (5.11) and (5.13) make it clear that the underlying model in the population cannot contain a full set of group/time interactions. So, for example, if the groups (cohorts) are defined by birth year, there cannot be a full set of birth year/time period interactions. We could allow this feature with invidual-level data because we would typically have variation in the covariates within each group/period cell. Thus, the absense of full cohort/time effects in the population model is the key an identifying restriction.

Even if we exclude full group/time effects, $\boldsymbol{\beta}$ may not be precisely estimable. Clearly $\boldsymbol{\beta}$ is not identified if we can write $\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g$ for vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_g$, $t = 1,\ldots,T$, $g = 1,\ldots,G$. In other words, while we must exclude a full set of group/time effects in the structural model, we need some interaction between them in the distribution of the covariates. One might be worried about this way of identifying $\boldsymbol{\beta}$. But even if we accept this identification strategy, the variation in $\{\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} : t = 1,..,T, g = 1,\ldots,G\}$ or $\{\Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} : t = 2,..,T, g = 1,\ldots,G\}$ might not be sufficient to learn much about $\boldsymbol{\beta}$ – even if we have pretty good estimates of the population means.

We are now ready to formally discuss estimation of $\boldsymbol{\beta}$. We have two formulas (and there are many more) that can be used directly, once we estimate the group/time means for y_t and $\mathbf{x}_t$. We can use either true panel data or repeated cross sections. Angrist (1991) used panel data and grouped the data by time period (after differencing). Our focus here is on the case where we do not have panel data, but the general discussion applies to either case. One difference is that, with independent cross sections, we need not account for dependence in the sample averages across *g* and *t* (except in the case of dynamic models).

Assume we have a random sample on $(\mathbf{x}_t, y_t)$ of size N_t, and we have specified the G groups or cohorts. Write $\{(\mathbf{x}_{it}, y_{it}) : i = 1,\ldots,N_t\}$. Some authors, wanting to avoid confusion with a true panel data set, prefer to replace *i* with $i(t)$ to emphasize that the cross section units are different in each time period. (Plus, several authors actually write the underlying model in terms of the pooled cross sections rather than using the underlying population model – a mistake, in my view.) As long as we understand that we have a random sample in each time period, and that random sample is used to estimate the group/time means, there should be no confusion.

For each random draw i, it is useful to let $\mathbf{r}_i = (r_{it1}, r_{it2}, \ldots, r_{itG})$ be a vector of group indicators, so $r_{itg} = 1$ if observation i is in group g. Then the sample average on the response variable in group/time cell (g,t) can be written as

$$\mu_{gt}^{y} = N_{gt}^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} = (N_{gt}/N_t)^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it}, \tag{5.14}$$

where $N_{gt} = \sum_{i=1}^{N_t} r_{itg}$ is properly treated as a random outcome. (This differs from standard stratified sampling, where the groups are first chosen and then random samples are obtained within each group (stratum). Here, we fix the groups and then randomly sample from the population, keeping track of the group for each draw.) Of course, μ_{gt}^{y} is generally consistent for μ_{gt}^{y}. First, $\hat{\rho}_{gt} = N_{gt}/N_t$ converges in probability to $\rho_g = P(r_{itg} = 1)$ – the fraction of the population in group or cohort g (which is supposed to be constant across t). So

$$\begin{aligned} &\hat{\rho}_{gt}^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} \xrightarrow{p} \rho_g^{-1} E(r_{itg} y_{it}) \\ &= \rho_g^{-1} [P(r_{itg} = 1) \cdot 0 + P(r_{itg} = 1) E(y_{it} | r_{itg} = 1)] \\ &= E(y_{it} | r_{itg} = 1) = \mu_{gt}^{y}. \end{aligned}$$

Naturally, the argument for other means is the same. Let $\mathbf{w}_{it}$ denote the $(K+1) \times 1$ vector $(y_{it}, \mathbf{x}_{it})'$. Then the asymptotic distribution of the full set of means is easy to obtain:

$$\sqrt{N_t}\,(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}} - \boldsymbol{\mu}_{gt}^{\mathbf{w}}) \rightarrow Normal(\mathbf{0}, \rho_g^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}),$$

where $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ is the sample average for group/time cell (g,t) and

$$\boldsymbol{\Omega}_{gt}^{\mathbf{w}} = Var(\mathbf{w}_t | g)$$

is the $(K+1) \times (K+1)$ variance matrix for group/time cell (g,t). When we stack the means across groups and time periods, it is helpful to have the result

$$\sqrt{N}\,(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}} - \boldsymbol{\mu}_{gt}^{\mathbf{w}}) \rightarrow Normal(\mathbf{0}, (\rho_g \kappa_t)^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}), \tag{5.15}$$

where $N = \sum_{t=1}^{T} N_t$ and $\kappa_t = \lim_{N \rightarrow \infty} (N_t/N)$ is, essentially, the fraction of all observations accounted for by cross section t. Of course, $\rho_g \kappa_t$ is consistently estimated by N_{gt}/N, and so, the implication of (5.15) is that the sample average for cell (g,t) gets weighted by N_{gt}/N, the fraction of all observations accounted for by cell (g,t).

In implementing minimum distance estimation, we need a consistent estimator of $\boldsymbol{\Omega}_{gt}^{\mathbf{w}}$, and the group/time sample variance serves that purpose:

$$\hat{\boldsymbol{\Omega}}_{gt}^{\mathbf{w}} = N_{gt}^{-1} \sum_{i=1} r_{itg}(\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})(\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})' \overset{p}{\to} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}. \tag{5.16}$$

Now let $\boldsymbol{\pi}$ be the vector of all cell means. For each (g,t), there are $K+1$ means, and so $\boldsymbol{\pi}$ is a $GT(K+1) \times 1$ vector. It makes sense to stack $\boldsymbol{\pi}$ starting with the $K+1$ means for $g = 1$, $t = 1$, $g = 1$, $t = 2$, ..., $g = 1$, $t = T$, ..., $g = G$, $t = 1$, ..., $g = G$, $t = T$. Now, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are always independent across g because we assume random sampling for each t. When $\mathbf{x}_t$ does not contain lags or leads, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are independent across t, too. (When we allow for lags of the response variable or explanatory variables, we will adjust the definition of $\boldsymbol{\pi}$ and the moment conditions. Thus, we will always assume that the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are independent across g and t.) Then,

$$\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \to \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}), \tag{5.17}$$

where $\boldsymbol{\Omega}$ is the $GT(K+1) \times GT(K+1)$ block diagonal matrix with (g,t) block $\boldsymbol{\Omega}_{gt}^{\mathbf{w}}/(\rho_g \kappa_t)$. Note that $\boldsymbol{\Omega}$ incorporates both different cell variance matrices as well as the different frequencies of observations.

The set of equations in (5.8) constitute the restrictions on $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$. Let $\boldsymbol{\theta}$ be the $(K+T+G-1)$ vector of these parameters, written as

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'.$$

There are $GT(K+1)$ restrictions in equations (5.8), so, in general, there are many overidentifying restrictions. We can write the set of equations in (5.8) as

$$\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}, \tag{5.18}$$

where $\mathbf{h}(\cdot,\cdot)$ is a $GT(K+1) \times 1$ vector. Because we have $\sqrt{N}$-asymptotically normal estimator $\hat{\boldsymbol{\pi}}$, a minimum distance approach suggests itself. It is different from the usual MD problem because the parameters do not appear in a separable way, but MD estimation is still possible. In fact, for the current application, $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is linear in each argument, which means MD estimators of $\boldsymbol{\theta}$ are in closed form.

Before obtaining the efficient MD estimator, we need, because of the nonseparability, an initial consistent estimator of $\boldsymbol{\theta}$. Probably the most straightforward is the "fixed effects" estimator described above, but where we estimate all components of $\boldsymbol{\theta}$. The estimator uses the just identified set of equations.

For notational simplicity, let $\boldsymbol{\mu}_{gt}$ denote the $(K+1) \times 1$ vector of group/time means for each (g,t) cell. Then let $\boldsymbol{\omega}_{gt}$ be the $(K+T+G-1) \times 1$ vector $(\boldsymbol{\mu}_{gt}^{\mathbf{x}}, d_t, c_g)'$, where d_t is a $1 \times (T-1)$ vector of time dummies and c_g is a $1 \times G$ vector of group dummies. Then the moment conditions are

$$\left(\sum_{g=1}^{G}\sum_{t=1}^{T} \boldsymbol{\omega}_{gt}\boldsymbol{\omega}_{gt}'\right)\boldsymbol{\theta} - \left(\sum_{g=1}^{G}\sum_{t=1}^{T} \boldsymbol{\omega}_{gt}\mu_{gt}^{y}\right) = \mathbf{0}. \tag{5.19}$$

When we plug in $\hat{\boldsymbol{\pi}}$ – that is, the sample averages for all (g,t), then $\check{\boldsymbol{\theta}}$ is obtained as the so-called "fixed effects" estimator with time and group effects. The equations can be written as

$$\mathbf{q}(\hat{\boldsymbol{\pi}}, \check{\boldsymbol{\theta}}) = \mathbf{0}, \tag{5.20}$$

and this representation can be used to find the asymptotic variance of $\sqrt{N}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta})$; naturally, it depends on $\boldsymbol{\Lambda}$ and is straightforward to estimate.

But there is a practically important point: there is nothing nonstandard about the MD problem, and bootstrapping is justified for obtaining asymptotic standard errors and test statistics. (Inoue (forthcoming) asserts that the "unconditional" limiting distribution of $\sqrt{N}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is not standard, but that is because he treats the sample means of the covariates and of the response variable differently; in effect, he conditions on the former.) The boostrapping is simple: resample each cross section separately, find the new groups for the bootstrap sample, and obtain the "fixed effects" estimates. It makes no sense here to resampling the groups.

Because of the nonlinear way that the covariate means appear in the estimation, the bootstrap may be preferred. The usual asymptotic normal approximation obtained from first-order asymptotics may not be especially good in this case, especially if $\sum_{g=1}^{G}\sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$ is close to being singular, in which case $\boldsymbol{\beta}$ is poorly identified. (Inoue (2007) provides evidence that the distribution of the "FE" estimator, and what he calls a GMM estimator that accounts for different cell sample sizes, do not appear to be normal even with fairly large cell sizes. But his setup for generating the data is different – in particular, he specifies equations directly for the repeated cross sections, and that is how he generates data. As mentioned above, his asymptotic analysis differ from the MD framework, and implies nonnormal limiting distributions. If the data are drawn for each cross section to satisfy the population panel data model, the cell sizes are reasonably large, and there is sufficient variation in $\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, the minimum distance estimators should have reasonable finite-sample properties. But because the limiting distribution depends on the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, which appear in a highly nonlinear way, asymptotic normal approximation might still be poor.)

With the restrictions written as in (5.18), Chamberlain (lecture notes) shows that the optimal weighting matrix is the inverse of

$$\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})', \tag{5.21}$$

where $\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is the $GT \times GT(K+1)$ Jacobian of $\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$. (In the standard case, $\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is the identity matrix.) We already have the consistent estimator of $\boldsymbol{\pi}$ – the cell averages – we showed how to consistently estimate $\boldsymbol{\Omega}$ in equations (5.16), and we can use $\check{\boldsymbol{\theta}}$ as the initial consistent estimator of $\boldsymbol{\theta}$.

$\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \nabla_{\pi}\mathbf{h}(\boldsymbol{\beta}) = \mathbf{I}_{GT} \otimes (-1,\boldsymbol{\beta}')$. Therefore, $\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is a block diagonal matrix with blocks

$$(-1,\boldsymbol{\beta}')(\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}_{gt}^{w}(-1,\boldsymbol{\beta}')'. \tag{5.22}$$

But

$$\tau_{gt}^2 \equiv (-1,\boldsymbol{\beta}')\boldsymbol{\Omega}_{gt}^{w}(-1,\boldsymbol{\beta}')' = Var(y_t - \mathbf{x}_t\boldsymbol{\beta}|g), \tag{5.23}$$

and a consistent estimator is simply

$$N_{gt}^{-1}\sum_{i=1}^{N_t} r_{itg}(y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - \check{\eta}_t - \check{\alpha}_g)^2$$

is the residual variance estimated within cell (g,t).

Now, $\nabla_{\theta}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \mathbf{W}(\boldsymbol{\pi})$, the $GT \times (K+T+G-1)$ matrix of “regressors” in the FE estimation, that is, the rows of $\mathbf{W}(\boldsymbol{\pi})$ are $\boldsymbol{\omega}_{gt} = (\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}, \mathbf{d}_t, \mathbf{c}_g)$. Now, the FOC for the optimal MD estimator is

$$\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}(\hat{\mathbf{W}}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{gt}^{y}) = \mathbf{0},$$

and so

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}_{gt}^{y}. \tag{5.24}$$

So, as in the standard cases, the efficient MD estimator looks like a “weighted least squares” estimator. The estimated asymptotic variance of $\hat{\boldsymbol{\theta}}$, following Chamberlain, is just $(\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}/N$. Because $\hat{\mathbf{C}}^{-1}$ is the diagonal matrix with entries $(N_{gt}/N)/\hat{\tau}_{gt}^2$, it is easy to weight each cell (g,t) and then compute both $\hat{\boldsymbol{\theta}}$ and its asymptotic standard errors via a weighted regression; fully efficient inference is straightforward. But one must compute the $\hat{\tau}_{gt}^{2}$ using the individual-level data in each group/time cell.

It is easily seen that the so-called "fixed effects" estimator, $\check{\boldsymbol{\theta}}$, is

$$\check{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\boldsymbol{\mu}}_{gt}^{y}, \tag{5.25}$$

that is, it uses the identity matrix as the weighting matrix. From Chamberlain (lecture notes), the asymptotic variance of $\check{\boldsymbol{\theta}}$ is estimated as $(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\check{\mathbf{C}}\hat{\mathbf{W}}(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}$, where $\check{\mathbf{C}}$ is the matrix described above but with $\check{\boldsymbol{\beta}}$ used to estimate the cell variances. (Note: This matrix cannot be computed by just using the "heteroskedasticity-robust" standard errors in the regress $\hat{\mu}_{gt}^{y}$ on $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, $\mathbf{d}_t$, $\mathbf{c}_g$.) Because inference using $\check{\boldsymbol{\theta}}$ requires calculating the group/time specific variances, we might as well use the efficient MD estimator in (5.24).

Of course, after the efficient MD estimation, we can readily compute the overidentifying restrictions, which would be rejected if the underlying model needs to include cohort/time effects in a richer fashion.

A few remaining comments are in order. First, several papers, including Deaton (1985), Verbeek and Nijman (1993), and Collado (1997), use a different asymptotic analysis. In the current notation, $GT \rightarrow \infty$ (Deaton) or $G \rightarrow \infty$, with the cell sizes fixed. These approaches seems unnatural for the way pseudo panels are constructed, and the thought experiment about how one might sample more and more groups is convoluted. While $T \rightarrow \infty$ conceptually makes sense, it is still the case that the available number of time periods is much smaller than the cross section sample sizes for each T. McKenzie (2004) has shown that estimators derived under large G asymptotics can have good properties under the MD asymptotics used here. One way to see this is that the IV estimators proposed by Collado (1997), Verbeek and Vella (2005), and others are just different ways of using the population moment conditions in (5.8).

(Some authors appear to want it both ways. For example, Verbeek and Nijman (1993) use large G asymptotics, but treat the within-cell variances and covariances as known. This stance assumes that one can get precise estimates of the second moments within each cell, which means that N_{gt} should be large.)

Basing estimation on (5.8) and using minimum ditance, assuming large cell sizes, makes application to models with lags relatively straightforward. The only difference now is that the vectors of means, $\{\boldsymbol{\mu}_{gt}^{\mathbf{w}} : g = 1,\ldots,G; t = 1,\ldots,T)$ now contain redundancies. (In other approaches to the problem, for example Collado (1997), McKenzie (2004), the problem with adding y_{t-1} to the population model is that it generates correlation in the estimating equation based on the pooled cross sections. Here, there is no conceptual distinction between having exogenous or endogenous elements in $\mathbf{x}_t$; all that matters is how adding one modifies the MD moment conditions. As an example, suppose we write

$$\begin{aligned} y_t &= \eta_t + \rho y_{t-1} + \mathbf{z}_t\boldsymbol{\gamma} + f + u_t \\ E(u_t|g) &= 0,\ g = 1,\ldots,G \end{aligned} \tag{5.26}$$

where g is the group number. Then (5.8) is still valid. But, now we would define the vector of means as $(\mu_{gt}^{y}, \boldsymbol{\mu}_{gt}^{\mathbf{z}})$, and appropriately pick off μ_{gt}^{y} in defining the moment conditions. The alternative is to define $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ to include $\mu_{g,t-1}^{y}$, but this results in a singularity in the asymptotic distribution of $\hat{\boldsymbol{\pi}}$. It is much more straightforward to keep only nonredundant elements in $\boldsymbol{\pi}$ and readjust how the moment conditions.are defined in terms of $\boldsymbol{\pi}$. When we take that approach, it becomes clear that we now have fewer moments to estimate the parameters. If $\mathbf{z}_t$ is $1 \times J$, we have now have $J + T + G$ parameters to estimate from $GT(J + 1)$ population moments. Still, we have added just one more parameter.

To the best of my knowledge, the treatment here is the first to follow the MD approach, applied to (5.8), to its logical conclusion. Its strength is that the estimation method is widely known and used, and it separates the underlyng population model from sampling assumptions. It also shows why we need not make any exogeneity assumptions on $\mathbf{x}_t$. Perhaps most importantly, it reveals the key identification condition: that separate group/time effects are not needed in the underlying model, but enough group/time variation in the means $E(\mathbf{x}_t|g)$ is needed to identify the structural parameters. This sort of condition falls out of other approaches to the problem, such as the instrumental variables approach of but it is harder to see. For example, Verbeek and Vella (2005) propose instrumental variables methods on the equation in time averages using interactions between group (cohort) and time dummies. With a full set of separate time and group effects in the main equation – derivable here from the population panel model – the key identification assumption is that a full set of group/time effects can be excluded from the structural equation, but the means of the covariates have to vary sufficiently across group/time. That is exactly the conclusion we reach with a minimum distance approach.

Interestingly, the MD approach applies easily to extensions of the basic model. For example, we can allow for unit-specific time trends (as in the random growth model of Heckman and Hotz (1989)):

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f_1 + f_2 t + u_t, \tag{5.27}$$

where, for a random draw i, the unobserved heterogeneity is of the form $f_{i1} + f_{i2}t$. Then, using the same arguments as before,

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g + \varphi_g t, \tag{5.28}$$

and this set of moment conditions is easily handled by extending the previous analysis. We can even estimate models with time-varying factor loads on the heterogeneity:

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + \lambda_t f + u_t,$$

where $\lambda_1 = 1$ (say) as a normalization. Now the population moments satisfy

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \lambda_t\alpha_g.$$

There are now $K + G + 2(T - 1)$ free parameters to estimate from $GT(K + 1)$ moments. This extension means that the estimating equations allow the group/time effects to enter more flexibly (although, of course, we cannot replace $\eta_t + \lambda_t\alpha_g$ with unrestricted group/time effects.) The MD estimation problem is now nonlinear because of the interaction term, $\lambda_t\alpha_g$. With more parameters and perhaps not much variation in the $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$, practical implementation may be a problem, but the theory is standard.

This literature would benefit from a careful simulation study, where data for each cross section are generated from the underlying population model, and where g_i – the group identifier – is randomly drawn, too. To be realistic, the underlying model should have full time effects. Verbeek and Vella (2005) come close, but they omit aggregate time effects in the main model while generating the explanatory variables to have means that differ by group/time cell. Probably this paints too optimistic a picture for how well the estimators can work in practice. Remember, even if we can get precise estimates of the cell means, the variation in $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ across g and t might not be enough to tie down $\boldsymbol{\beta}$ precisely.

Finally, we can come back to the comment about how the moment conditions in (5.8) only use the assumption $E(u_t|g) = 0$ for all t and g. It seems likely that we should be able to exploit contemporaneous exogeneity assumptions. Let $\mathbf{z}_t$ be a set of observed variables such that $E(u_t|\mathbf{z}_t, f) = \mathbf{0}$, $t = 1,\ldots,T$. (In a true panel, these vary across i and t. We might have $\mathbf{z}_t = \mathbf{x}_t$, but perhaps $\mathbf{z}_t$ is just a subset of $\mathbf{x}_t$, or we have extra instruments.) Then we can add to (5.8) the moment conditions

$$\begin{aligned} E(\mathbf{z}_t' y_t|g) &= \eta_t E(\mathbf{z}_t|g) + E(\mathbf{z}_t'\mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t' f|g) + E(\mathbf{z}_t' u_t|g) \\ &= \eta_t E(\mathbf{z}_t|g) + E(\mathbf{z}_t'\mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t' f|g), \end{aligned} \tag{5.29}$$

where $E(\mathbf{z}_t' u_t|g) = \mathbf{0}$ when we view group designation as contained in f. The moments $E(\mathbf{z}_t' y_t|g)$, $E(\mathbf{z}_t|g)$, and $E(\mathbf{z}_t'\mathbf{x}_t|g)$ can all be estimated by random samples from each cross section, where we average within group/time period. (This would not work if $\mathbf{x}_t$ or $\mathbf{z}_t$ contains lags.) This would appear to add many more moment restrictions that should be useful for identifying $\boldsymbol{\beta}$, but that depends on what we assume about the unobserved moments $E(\mathbf{z}_t' f|g)$.

References

(To be added.)